International Conference

Big Data, Knowledge and Control Systems Engineering

BdKCSE'2015

Sofia, Bulgaria
5-6 November 2015

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'2015

5-6 November 2015
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

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’2015) 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’2015 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

Plenary session „Big Data Foundations”

1. **Dr. Anton Gerunov** „Open Data: Police and Implementation in Bulgaria” 1

Invited Research/Industry Presentations „Big Data Management”

2. **Igor Mishkovski, Sanja Šćepanović, Ivan Klandev** „Creating Resilience Using Twitter Data" 13
3. **Elisavet Charalambous, Nectarios Efstathiou, Nikolaos Koutras** „A Cost Effective Solution for Audio Surveillance Using Embedded Devices as Part of a Cloud Infrastructure” 21
4. **Georgi Pavlov, Pavel Genevski** „Semantic Technologies for the Masses” 33

Paper session „Big Data Applications”

5. **Elisavet Charalambous, Romaios Bratskas, George Karkas, Andreas Anastasiades, Nikolaos Koutras** „An innovative Digital Forensic Tool assisting evidence analysis in Cyprus” 45
6. **Jordan Jordanov, Bogdan Vatkov, Dimitar Angelov** „Advanced Analytics with SAP HANA” 55
7. **Vasil Kolev** „Big Data in Astroinformatics - Compression of Scanned Astronomical Photographic Plates" 65
8. **Tanya Dilkovska-Petrova, Anna Zagorska, Kristina Bliznakova** „The Role of Cloud Services in Research Projects (Case Study)” 75
9. **Valentina Terzieva, Katia Todorova, Petia Kademova-Katzarova** „Big Data for Personalization: Opportunities, Applications” 89

Plenary session „Intelligent control systems and Data Management”

10. **Rumen Nikolov, Boyan Jekov, Petya Mihaylova** „Big Data in a Smart City Ecosystem: Models, Challenges and Trends” 101
11. **Mariyana Lyubenova, Yuri Pavlov, Alexandre Chikalanov, Kamen Spassov** „A Decision Making Support Model for a Smart Forest Ecological Management System" 119
12. **Roumen Nikolov, Elena Shoikova, Alexander Chikalanov, Galia Novakova** „Bike Sharing System for Multimodal Mobility Based on Open Data" 133
13. **František Čapkovič, Lyubka Doukovska, Vassia Atanassova** „Automated Control Synthesis for Discrete Event Systems" 139
14. **Nina Dobrinkova, Valentina Terzieva, Katia Todorova** „Knowledge and Data Management in ‘Civil Protection for ALL’ project" 149

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
Peđa 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

Open Data: Policy and Implementation in Bulgaria

Anton Gerunov
Council of Ministers of the Republic of Bulgaria and Sofia University "St. Kliment Ohridski"
125 Tsarigradsko Shosse Blvd, 1113 Sofia, Bulgaria
gerunov@uni-sofia.bg

Abstract: This paper provides an overview of Bulgaria's policy on open data by using the framework of a novel Open Data stage model. We also present a detailed overview the data landscape of the public sector in the country and outline the first milestones in its efforts to publish some of that data for reuse. The paper outlines the limited knowledge about true open data impact and proposes a few approaches to measure it formally in order to critically assess the usefulness of opening data.

Keywords: Open data, OGD, data audit, open data portal

1 Introduction

Open Data has received enormous attention as an integral part of the Open Government Movement. The concept revolves around the need of public bodies to share the information they collect in the regular course of their activities with the public so that this data can be reused and unlock additional economic and social value [1]. Open data encompasses much more than open government data (OGD) alone but the public sector seems to be the frontrunner in opening its data at this stage, followed closely by the scientific community in opening research data and promoting open science [2]. The private and the non-profit sector will likely be more proactive over the next years.

The main driving forces behind opening data concentrate around three major benefits that it is supposed to bring to the economy and the public space. These are the following [1], [3], [4]:

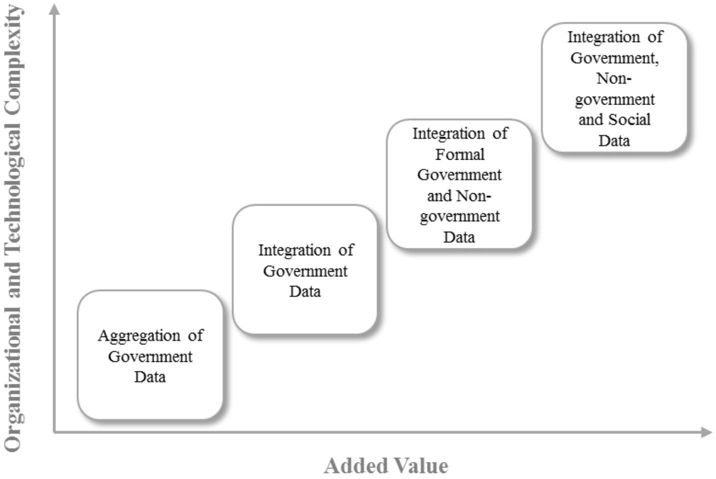
- Unlocking economic value by stimulating private sector innovation, providing for better business decision-making, and promoting the data science industry.
- Providing political and social benefits by increasing the public's capacity to monitor government and participate more effectively in the formulation of public policy.
- Realizing operation improvement in the public sector by removing duplications, improving the access to information and optimizing processes through benchmarking and reengineering.

Those potential benefits have led many countries to devise and implement open data policies in the hope of unlocking its large potential, and Bulgaria stands to be one of those countries. This paper will review the country's efforts within the framework of an Open Data stage model in Section 2 and report on the data landscape of the public sector in Section 3. It will also explore the impact of open data initiatives in Section 4, and give policy recommendations in Section 5. Section 6 concludes.

2 A Stage Model of Open Data

The theory and practice of OGD is relatively new and therefore few theoretical models are available to put a conceptual framework around a country's efforts. One can borrow from neighboring fields such as electronic government [5] but these are not fully applicable. For the purposes of this paper we are going to use a specific Open Data Stage model, developed by Kalampokis et al. [6]. This model is based on four main stages of opening data, each with increased sophistication and difficulty, but also with increased potential of unlocking value. The model is presented graphically in Figure 1.

Figure 1: A Stage Model of Open Data, Kalampokis et al.


Organizational and Technological Complexity
Aggregation of Government Data
Integration of Government Data
Integration of Formal Government and Non-government Data
Integration of Government, Non-government and Social Data
Added Value

The first stage consists of aggregating government data, whereby public sector agencies and units create inventories of their data, export it, and publish it. The format used is usually a well-known machine-readable format such as csv, json, xml, and others. The first stage is crucial as it has to overcome a number of technological, legislative, economic, and cultural barriers as many organizations are both unable and sometimes unwilling to freely share their data. To achieve success during this stage it is crucial to modify the data sharing culture and ensure sustainability for the overall process. There is some economic value to be unlocked by this stage as the public and private companies have access to previously unutilized data.

The second stage takes the data sharing efforts a step further by providing for integration of OGD contained in different databases. Thus it turns information into knowledge by creating a unified data view focused on a given object (company, region, government entity, etc.) by connecting all data in the public sector that pertains to it. This stage faces significant technological and organizational barriers. On the technological side, data formats must be compatible with each other, so that integration is possible. Currently, the Linked Open Data paradigm seems to be a leading contender to achieve this. On the organizational side challenges are even bigger. Data integration will likely show duplicated data and will reveal errors and inconsistencies in the information the public sector uses for decision-making. This may lead to erosion of public trust and inspire increased corrective action – two thing public sector bodies would rather avoid. We should note that initially integration is likely to be only partial and only as time passes it will be increasingly complete.

The third stage seeks to create even further opportunities by integrating formal government data with formal non-government data coming from the private sector, the media, or civil society organizations. A lot of non-public entities collect and maintain databases that can be potentially useful as they provide further knowledge for a given object of interest. Integrating OGD with those provides for even greater economic value. This stage is particularly challenging as it is the first that requires concerted efforts outside the public sector and thus outside the reach of government executive authority. Private companies and NGOs need to be convinced to share their data and it must also be processed so that integration with existing OGD is possible. Further, this pursuit may not be fruitful for all sorts of data since the use cases are not always clear and even in the event of positive economic benefit the implementation costs may be prohibitively high.

The fourth and final stage provides for integration of formal government and non-government data with social data, including information from social networks. Social data is created and voluntarily shared by citizens and often expresses opinion, belief, attitudes, and values. A large quantity of this data is streamed live through networks such as Facebook and Twitter and is sometimes accessible through their APIs. Such data integration can make for very sophisticated queries and can serve to spur innovation in economic production and government. This, however, comes at the prices of very high implementation costs and raises questions about privacy and control. The latter two stages – three and four – at this point seem difficult to achieve at a large scale but some instances of individuals and companies combining data from different sources abound.

Using the Stage Model as a framework, one can classify Bulgaria as belonging to the first stage, making initial steps with data inventories and beginning to publish key datasets of public interest on its Open Data Portal. At this initial stage the country faces its associated challenges – lack of data overview, difficulty in standardizing and publishing data, reluctance of public bodies to share information, and still limited use cases. In the next section we present the results of a data audit in the country to outline the starting position of OGD initiatives and then outline the concrete steps Bulgaria has taken to open data and unlock their value.

3 Data Resources in Bulgaria

Knowledge of the data landscape in the public sector is crucial for an open data initiative to succeed. This is sometimes challenging, as the government is not a monolithic entity but rather a collection of functionally organized administrations with set goals and related data they collect for their purposes. There are 576 administrative units that offer services to citizens, as registered in the Administrative Registry. Additionally, there are more than 2,363 different administrative services provided by these units, and most of them have informational requirements [7]. Data collection and access is either mandated by law, dictated by practical necessity, or done for historical reasons. This section outlines the key results from a full information audit of the public administration in Bulgaria and shows how it informed OGD policy and implementation.

3.1 Key Results from the Data Audit in Bulgaria

The preliminary data audit was done over the last quarter of 2014 and the first quarter of 2015 by requesting a complete data questionnaire from every single administrative unit. Since this was an initiative led by the Council of Ministers it had a very high response rate well over 90%, or a total of 564 responses received. In addition to that further research was undertaken to add more informational sources mandated by in order to complete the list. The audit shows a total of 8,156 different data sources in the administrations that are kept in over 1,300 server spaces and numerous work stations. The primary way of storing public sector data is via means of an internal server with 31% of respondents mentioning this. This is closely followed by storing data on external drives, paper and other means (30%), and on local workstations (25%). Outsourcing data storage remains unpopular with only 10% of administrative units using external servers or hosting. This indicates a culture of reluctance to disclose data and preference for internal handling and may be interpreted as symptomatic of a reserved attitude towards opening data.

Figure 2: Storage of Data in the Public Sector in Bulgaria

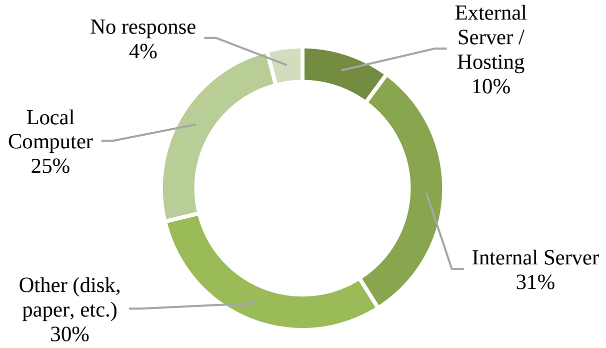


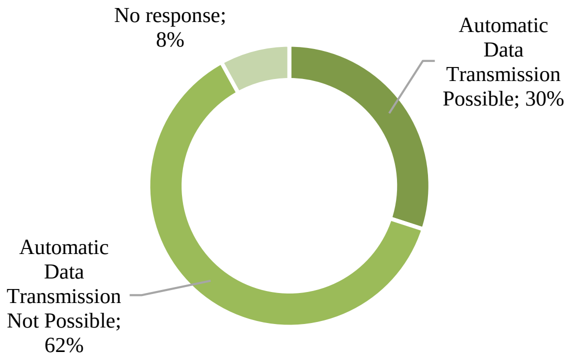
Table 1 presents an overview of the formats, used by administrations. The first striking conclusion is that an overwhelming amount of data is stored in formats that are difficult to process for further analysis. Word files amount to 9.1% of all data, pdf files – for another 4.1%, and fully 29.3% of all data is still stored exclusively on paper. This includes not only internal and external documents but also registries that are used in the process of service delivery. Structured formats also feature prominently, with almost 22% of information stored in Excel files. Fully machine readable files, that are ready for further processing include .mdb files (4.4%), .html files (4.1%), .xml and other databases (1.5%). About 19% of administrations could not exactly specify their data storage and opted for the answer "Other".

Table 1: Data formats used in the public sector

.doc/ .docx	**.html**	**.mdb**	**.pdf**	**.xls/ .xlsx**	**.xml**	**Data Base**	**Paper**	**Other**	**No reponse**	**Total**
741	334	362	338	1786	76	49	2396	1589	484	8171
9.1%	4.1%	4.4%	4.1%	21.9%	0.9%	0.6%	29.3%	19.6%	5.9%	100.0%

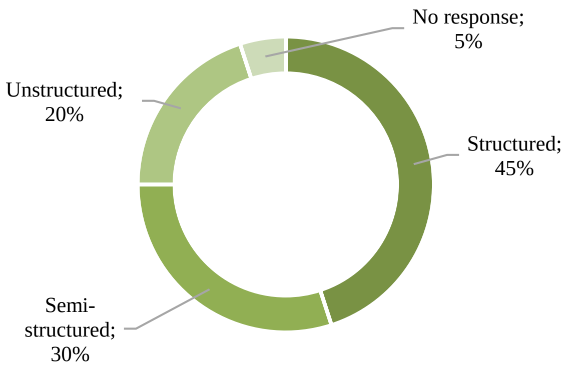
It is of particular note that information is still stored predominantly on paper, or in text formats. Even formats that should be machine-readable or easily export to a machine-readable form are often structured in such a way to hinder the process. This holds particularly true for Excel files, which are sometimes formatted in view more of their visual appeal and less in terms of interoperability (including merged cells, inserted columns and rows for formatting purpose, using visual templates, etc.) In total, this makes for a large number of files that are not truly machine-readable – 46% of registries are not, while only 41% of them are. For the rest, no response was given. Automatic data access is crucial for the provision of services and for the improvement of governance processes. It is also no less important for the OGD initiative as it makes for easier extraction and updates. Most of the data in the public sector, however, cannot be accessed automatically (Fig. 3).

Figure 3: Automatic Data Transmission



The overall conclusion is that less than half of the data audited consists of structured data – be it textual or numeric. There is a large chunk of semi-structured data, and about one fifth is completely unstructured (Fig. 3). This calls for significant efforts at structuring and standardizing data before opening and sharing it with the public. The process is likely to be hindered by the wide unavailability of meta-data – fully 64% of the respondents say that no metadata is available, and only 28% of public sector information is accompanied by some metadata.

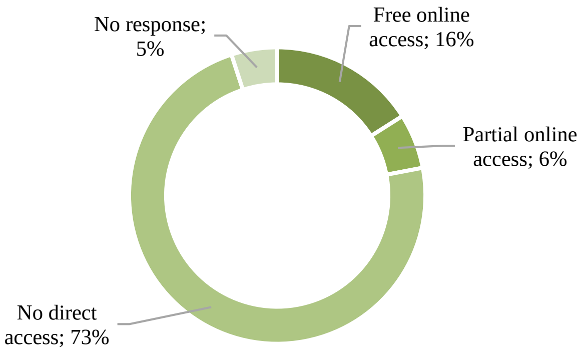
Figure 4: Structure of the data in the public administration



A final question of key interest for the OGD is whether any of this data is accessible to the public. If data is already published, its aggregation on a single access point brings the benefit of convenience and increases awareness. If the data has never been publicly available, its publication has the potential to unlock both social and economic benefits as actors in the public domain experience a tangible increase in disposable information.

A small fraction of public sector information is available to the public – only 16% of data has free online access. An additional 6% of data is published only in part at the discretion of the administration. Fully 73% of data sources remain unavailable for reuse by private agents and this comprises a measure of OGD's potential to generate business and social innovation. The large share of undisclosed information clearly outlines the need for a targeted and comprehensive open data policy that will ensure effective reuse of data and provide for sustainability of this process.

Figure 5: Data availability for public access



3.2 Milestones in Opening Data

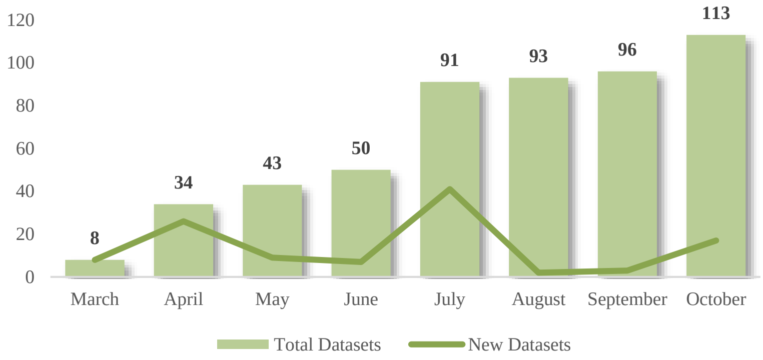
Bulgaria has become active in the process of opening public sector data relatively late in respect to other EU countries. In terms of legislation, the grounds for publishing OGD at the EU level is given in Directive 2013/37/EC, which is transposed in Bulgarian legislation by means of modification of the Freedom to Information Act. A demonstration portal for open data is active since October 2014 and policy efforts to populate it have been undertaken since then. National OGD policy has been enshrined as country reform plans in pursuit of Bulgaria's second Open Government Action plan under the international initiative Open Government Partnership (OGP). In it the country pledged to publish extensive public sector information. Initially the list of data to be published was prioritized by an Open Data task force in the Council of Ministers.

The task force gave particular weight to the relative importance of data, its ability to disclose government expenditure, to inform policy debate on issues such as healthcare, education, environment, law enforcement, and to aid in improving public sector efficiency.

Thus the task force came up with a list of 119 priority datasets to be opened that included the public procurement register and budget payments, educational performance statistics, air pollution data, vacant jobs data, road accidents and crime statistics, business and demographic statistics, donors in kind to political parties, etc. The list was adopted by the Council of Ministers with CM Act 103/2015 and public sector organization were required to open those datasets and publish them on the single point of access – the online portal *opendata.government.bg*.

Following the Council for Administrative Reform schedule, public sector units started uploading data since March 2015. Progress has been significant (Fig. 6) and the total number of datasets by the end of 2015 will probably go beyond the initially planned 119. Further prioritization is also possible through the feedback form on the portal that allows users to request data of interest to them.

Figure 6: Number of available datasets on the open data portal, Q1-Q3.2015



In conjunction with the top down approach, the government has mounted initiatives to popularize the use and reuse of OGD. Among those are press conferences and seminars such as "A Date with Data" in July 2015, and a number of trainings for about 100 data administrators.

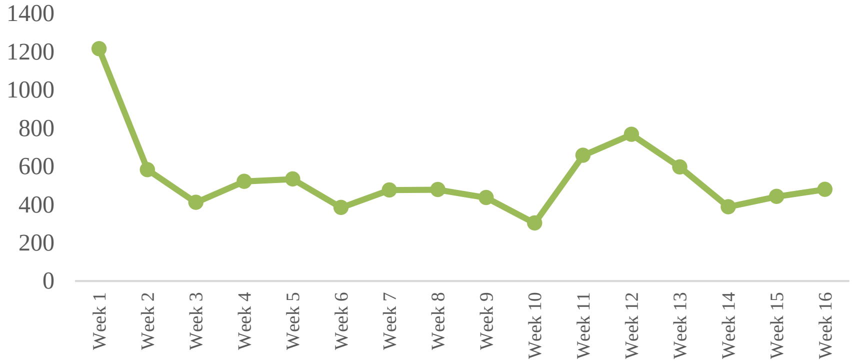
4 Open Data Impact

The impact of opening up data is often debated and espoused as the primary reason for publishing OGD [1], [3]. While recourse to its economic and democratic impact is seen as useful driver for publicizing more data it is rarely easy to quantify the impact this initiative has on business and society. So far, efforts at measuring impact has been mixed and unable to produce concrete results on the usefulness of OGD [8]. The crux of the issues lies in the fact that merely opening up datasets does not automatically mean that the public can use them meaningfully or that business can profitably utilize them. Apart from access, the impact of open data depends crucially on engagement, ability to analyze and draw conclusions from information, and a suitable institutional and economic environment that is receptive of such innovation. In fact, barriers to usage of open data are sometimes seen as so high that some authors argue that OGD empowers the already empowered – the highly educated persons and sophisticated businesses that can extract value from public information [9]. All this is likely to put real world open data impact as smaller and more unequal than usually discussed in public policy circles.

4.1 Usage Metrics and Anecdotal Evidence

A common approach for measuring policy success is through the sheer number of available datasets as this is supposed to capture the availability of public data for reuse. As of the beginning of November 2015, Bulgaria has published a total of 113 datasets on its single point of access and continues towards the target of 119. Such a measure misrepresents the impact of data as it fails to discern both the public's exposure to information, as well as its quality and usefulness. How often users leverage OGD can easily be measured by a metric of the open data portal's popularity such as visits, unique users, or average session duration. By this measure open datasets in Bulgaria score relatively well with an average monthly visits of 2,167 (Fig. 7), but still some improvement is possible. For a period of 4 months from 15 July 2015 to 31 October, the site has had 8,668 sessions and 6,188 users. Of those 72.3% were new visitors, while the rest were return visitors.

Figure 7: Active Session on Open Data Portal, 15.07.2015-31.10.2015



The quality, usefulness and impact of information is harder to gauge. While there is some anecdotal evidence of effective data usage, total impact remains elusive. Despite this, a few success stories have proliferated giving examples of OGD potential. One of them is connected with the release of new voter registration in the run-up to the October 2015 municipal elections which showed a huge abnormal spike in registrations and propelled authorities to take action against it. While informative such case studies need to be supplemented by a fuller and more comprehensive account of OGD's usefulness.

4.2 True Impact Metrics

Impact metrics need to quantify both economic and political benefits brought about by the totality of open data, and also take account of the distribution of those benefits. We can outline two major approaches to measuring this impact:

- **Macro-level approaches** – since OGD is supposed to stimulate information and improve the public environment, it should be the case that it is associated with a measure of technological development such as Total Factor Productivity (TFP). A possible approach is to use a general linear model with TFP as dependent variable, and a measure of OGD and a vector of controls as independent variables. While imperfect, this approach can give an estimate whether OGD has a transformative power for the overall economy and server as a useful guide for policy discussion.
- **Meso-level approaches** – opening data in a specific sector should bring notable improvement in it, which can be seen in some pre-determined data indicators. For example, opening procurement data should lead to more transparency and less corruption and thus lower the price for reference orders. Other causal effects can be ruled out or controlled for using the GLM model outlined above. Such metrics have the power to provide the very concrete benefits for opening data and will be especially useful for improving the efficiency and effectiveness of public sector units and organizations.
- **Micro-level** approaches – these focus on a specific datasets or groups of datasets, and follow them through their lifecycle. By doing this, the researcher gets a full and nuanced picture of usage, impact, and benefit distribution. The most common micro-level approach is the case study whereby each OGD dataset usage is described in detail, giving the context and measuring benefits to different stakeholders [10]. Case studies generally use a mixed methods design and serve as an excellent illustration of OGD potential. They can thus be leveraged as a powerful argument in favor of openness. The main issues with this approach is that it fails to scale well and is suffering from observer bias.

The method of choice for measuring impact naturally differs across situations and has to adapt to the context of specific data openness. What is most important is not to overlook this key aspect of OGD policy. It is indeed difficult to manage something that is not measured.

5 Conclusions

The paper makes an overview of Bulgaria's open data policy and puts it in the context of a stage model to outline possible future directions. For a very short period of time the country has disclosed a relatively large number of datasets with large economic and social potential. While data usage and popularization is still in need of improvement, there is already some anecdotal evidence of OGD's impact.

Despite the short period of focused policy, Bulgaria has reached a level of open data maturity which allows for a number of important further developments. First, it needs to spell out a specific Open Data strategy outside the framework of OGP to underline its political commitment. Second, more targeted efforts at measuring OGD impact will serve as important drivers for sustainability and expansion of the policy efforts. Third, and most important, the country needs to make more confident strides to the second stage of OGD development by linking government datasets in order to unlock even more value. In conclusion, if fast-paced progress is maintained, OGD has the potential to serve as a transformative power for the public and private sectors alike.

References

[1] Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4), 258-268.

[2] Molloy, J. C. (2011). The open knowledge foundation: open data means better science. *PLoS Biol*, 9(12), e1001195.

[3] Huijboom, N., & Van den Broek, T. (2011). Open data: an international comparison of strategies. *European journal of ePractice*, 12(1), 4-16.

[4] Davies, T. (2010). *Open data, democracy and public sector reform. A look at open government data use from data.gov.uk*. UK: Open Data Impacts.

[5] Directorate General for Information Society and Media. (2009). *Smarter, Faster, Better eGovernment. 8th eGovernment Benchmark Measurement*. Brussels: European Commission.

[6] Kalampokis, E., Tambouris, E., & Tarabanis, K. (2011). Open government data: A stage model. *Electronic government* (pp. 235-246). Springer Berlin Heidelberg.

[7] Council for Administrative Reform. (2015). *Report on the Public Administration, 2014*. Sofia: Council of Ministers.

[8] Bertot, J. C., McDermott, P., & Smith, T. (2012). Measurement of open government: Metrics and process. In System Science (HICSS), *45th Hawaii International Conference on* (pp. 2491-2499). IEEE.

[9] Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone?. *First Monday*, 16(2).

[10] GovLab. (2015). *Open Data Impact Case Studies*. New York University: The Government Lab.

Creating Resilience Using Twitter Data

Igor Mishkovski[1], Sanja Šćepanović[2], Ivan Klandev[1]

[1]Faculty of Computer Science and Engineering, Ss. Cyril and Methodius, Skopje, R. Macedonia

igor.mishkovski@finki.ukim.mk

i.klandev@gmail.com

[2]Aalto University School of Science, Helsinki, Finland

sanja.scepanovic@aalto.fi

Abstract: Analysing people's behaviours on social networks and their communication patterns provides insight into human habits and their typical actions on different timescales. Combining these results with data available from other important sectors enables the study of the interrelationship and influence of major external events on the changes in the "normal" communication and social pattern behaviours. Towards this goal, we have developed a platform that uses twitter stream in order to process crisis-related communication. The developed platform was tested for earthquakes and floods in the Balkan region in the period between April and June 2014, when the largest recorder floods in the Balkans region happened.

Keywords: Big Data, Twitter, DRR, Resilience

1 Introduction

Human behaviour before, during and after certain catastrophic event is a hot topic in network and social sciences. Moreover, the availability of (big) data that carry location and may carry context information: from the traveling bank-notes [1], mobile phone logs [2][3][4] or traveling smart cards [5], to the on-line check-ins and statuses with geo-locations obtained from social networks, such as Foursquare [6] and Twitter [7], boost the researchers' contribution in the field of Disaster Risk Reduction (DRR) and in creating resilience for catastrophic events.

The location-based data associated with a given context, as in Twitter, has been proven useful for practical studies in several important sectors; for example, for modelling and future planning, and real-time and aftermath analysis in disaster response and disaster risk reduction, and for health, socio-economic and transportation sector. A potential is also recognized for a planet-scale mobility measurement [8] through opening to researchers and combining the different big datasets carrying location information.

Some examples of using big data for creating resilience and DRR are given in the following. The authors in [9] show the potential for urban studies and planning by using location-based services, in particular from mobile phone datasets from the city of Milano. Authors in [10] provided partial solution to the traffic congestion in the city of Abidjan using the CDR dataset for Côte d'Ivoire. Their analysis shows that by adding four additional routes and extending one existing route the people in Abidjan will reduce their travel time by 10%. Another field that has great benefit from time-sensitive information obtained from location based data is the disaster response and disaster risk reduction, in terms of fast resource allocation and directing the emergency aid, as well as in analysing the people's movement and migration after some emergency situation [11]. The results in [12] emphasized that social bounds are crucial for people's movement in case of earthquake. This kind of analysis could be further used to estimate the post-catastrophic population situation in a given region and to plan emergency aid more efficiently. In [2] the authors show how emergency situations can be detected by only observing normal collective calling patterns and alerting those patterns that exceed threshold around mean activity.

Today, especially in developing countries, the process of obtaining relevant real-world indicators is a hard task that needs a lot of expertise and resources. Thus, it is desirable if the results from social network data analysis can be used as proxy indicators to estimate and give insights of the socioeconomic situation of one country. For instance, it has been shown that the diversity of individuals' relationships can be used as an indicator of the economic development of certain communities. The more the diverse they are the better the access exist to social and economic opportunities [13]. Another research in Latin America showed that the reciprocity of communications, the physical distance with the contacts and the area in which people move is tightly connected to the socio-economic level of a person and the expenses [14]. In [15] authors have detected and validated the poverty levels of the 11 different regions. Thus, they again prove that the provided data can be used as a proxy indicator for assessing the health, education, living standard and the threat from violence.

Recently, microblogs, such as Twitter has been shown that can have great impact in emergency situations [16] of natural disasters such as: earthquakes [17], floods [18], hurricanes [19], and wildfires [20]. Social media's technology platforms allow for multidirectional network communication which can aid officials during disasters to compile a list of the injured, deceased, and contact family and friends of victims [21]. This provides public and mental health value to the population affected by connecting vital services and resources [22][23].

Finally, sometimes the social microblog platform can be used as a better early warning system, compared to the traditional media [24].

In this work, we present the platform makes temporal-spatial-context analyses of the Twitter user data in order to understand the users' behaviour when emergency situations and/or natural disasters occur. Moreover, we have tested the platform on the data twitted when the biggest flood happened in the Balkan region, presenting several key points: i) sometimes the centralized government response can be misleading or questionable compared to the twitter data analysis, ii) the role of influencers in the social microblog platform is crucial for DRR response and iii) technological platforms based on social microblogging can be used as an early warning system for a certain type of disaster events.

The structure of this paper is as follow. The platform is presented in Section 2 whereas the results from the analysis of the Twitter data connected to the Balkan floods are given in Section 3. Section 4 concludes this work.

2 A platform for Twitter DRR data analysis

In this section we present a Web platform that uses the Twitter stream in order to process crisis-related communication, see [25]. The DRR platform in Figure 1 captures tweets which by their content belong to some DRR category and can be used for different kind of after-math reports and crisis analysis. For instance, the table data view, shows the most important properties of relevant tweets, such as: timestamp, tweets content, number of retweets and the category to which the tweet belongs. The DRR stakeholders can decide if a given tweet is valid for a certain category. In this way, we plan to introduce an intelligence to the system, i.e. the system can learn which tweets to be taken into account as valid tweets, according to the tweeter account, the tweet text, etc.

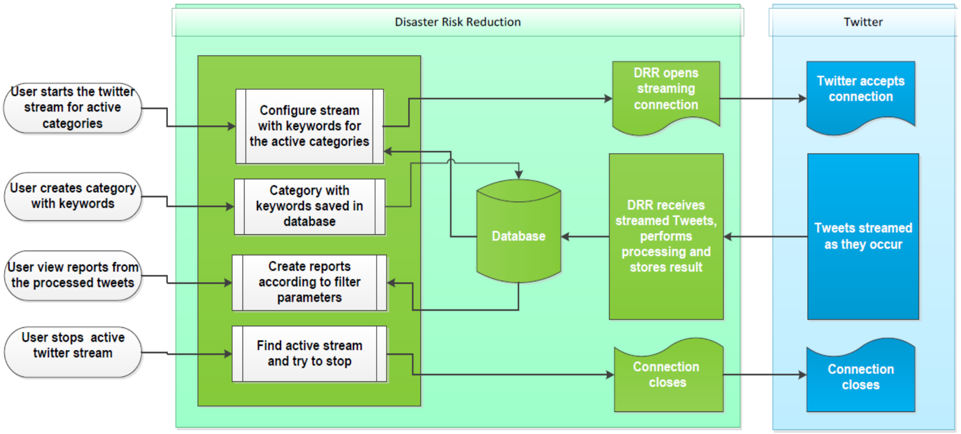

Disaster Risk Reduction
Twitter
User starts the twitter stream for active categories
User creates category with keywords
User view reports from the processed tweets
User stops active twitter stream
Configure stream with keywords for the active categories
Category with keywords saved in database
Create reports according to filter parameters
Find active stream and try to stop
Database
DRR opens streaming connection
DRR receives streamed Tweets, performs processing and stores result
Connection closes
Twitter accepts connection
Tweets streamed as they occur
Connection closes

Figure 1 Twitter DRR Platform

The platform will trigger a possible DRR event, if there is an increased activity on a given DRR category using the trend-line chart. Moreover, a hashtag cloud is another type of report, useful to find out the most used hashtags for a given category. Finally, the platform visualize the tweets according to the geo-coordinates. Here, we are using two ways to geo-map the tweets: i) using the GPS coordinates provided by the tweets and ii) using the tweet text and finding the geographical names in the tweets (using geonames.org) and then we use reverse geo-coding to obtain the coordinates of the place in the tweet text. The later, could be useful when the GPS coordinates are not shared in the tweet and when people from one place tweet about possible disaster (such as floods) that might occur in other place.

3 DRR results obtained from the platform

The platform was tested for earthquakes and floods in period of several months in 2014. For the earthquake that happened on 25.05.2014 at 23:05:00 in Skopje, the DDR platform alerted on 25.05.2014 at 23:06:03, whereas the first official information came on 25.05.2014 at 23:16:00. Thus, it is evident that DRR tweets can be used as an early warning for some DRR events, such as earthquakes. Moreover, in the following we present some interesting results about the floods in Serbia in May 2014. This example presents another type of use-case for our platform, namely, for post-disaster analysis and raising global and fast awareness about the crisis after the disaster.

The floods in the Balkans area during May 2014 were the largest in recorded history. Starting on 13th of May, culminating on 15th and 16th, and subsiding in different places from 18th to 20th of May, the floods have affected over 1.6 million people with estimated damage only in Serbia over 1.5 billion euros. Considering the slow nature of flood disaster, tweets have not shown to be useful as an early warning: on 13th, we still did not find many tweets with the keyword "poplava" ("flood" in local languages). However, a more detailed analysis on 15th shows exceptionally high number of tweets containing geo location Ub (a town and municipality located near the river Kolubara), see Fig. 2. On the night between 15th and 16th the dam on the river Kolubara broke and that resulted with catastrophic consequences for the town of Obrenovac. Additional social network data revealed that people were aware of this potential catastrophe and have been warning about it online before it happened. The centralized governmental channels, however, advised people not to leave their homes [26]. Thus, in the case of Serbia, a potential of crowd-sourced disaster warning has been confirmed with our DRR platform, however, this time, the affected people still listened to the centralized government news, what did not lead to an optimal disaster response.

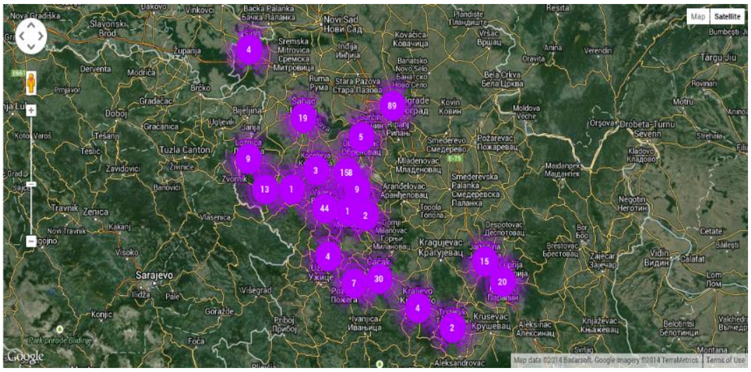


Figure 2 Number of flood related tweets for affected geo-locations on 15th of May

On Saturday the 17th at 20:00 CET we find an unusual spike: the hourly number of captured tweets almost tripled (Fig. 3). Our analysis shows that such a spike points out the power of influencers when it comes to the disaster information spreading.

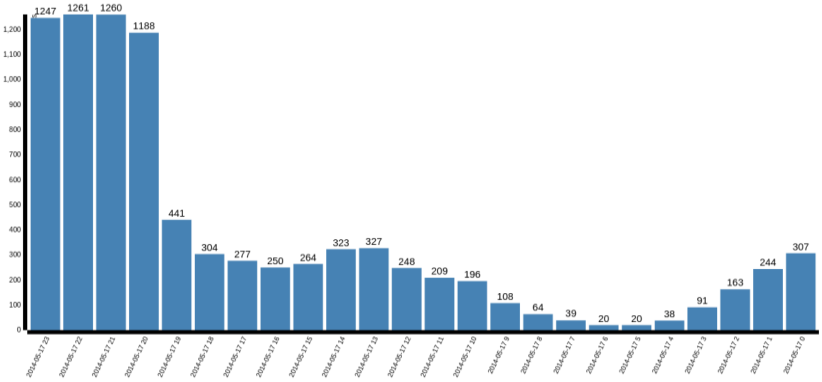


Figure 3 Hourly number of flood related tweets on 17th of May

Namely, the spike happens following the tweet of the wife of Serbian tennis player Novak Djokovic. She tweeted a picture of him holding a transparent with Boris Becker and sending support for Serbia after winning the semi-finals of the tournament in Rome. In addition to the change in the number of tweets, the social DRR platform presented us with the important change in geo-locations from which tweets come from, after Djokovic's intervention (see Figure 4), as well as in the hashtag cloud content (previously dominating were mainly descritpitve hashtags: #poplava and #SerbiaFloods, while afterwards, the most dominant is a more action-oriented hashtag #HelpForSerbia). Djokovic also criticised BBC, CNN and other largest publishers for not reporting enough about the floods. Irrespective of whether we would agree with Djokovic's critique of the publishers, our analysis shows that it was effective in a way that after it the news about floods have spread on Twitter, and on the news media [27] around the world.

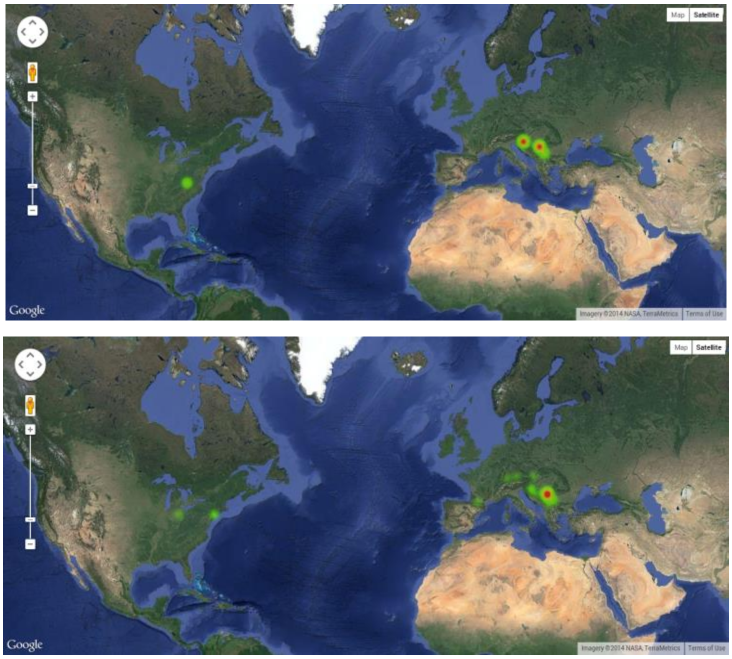

Map
Satellite
Google
Imagery ©2014 NASA, TerraMetrics
Terms of Use
Map
Satellite
Google
Imagery ©2014 NASA, TerraMetrics
Terms of Use

Figure 4 Thermal map with locations of flood related tweets on 17th of May at 19h (above) and 20h (bellow).

4 Conclusions

In summary, our analysis of the Serbian floods using the DRR platform, results with few important points: the wisdom of the crowd can be more effective compared to the centralized governmental response, the power of online influencers can be employed in DRR response and that for some disaster events, such as earthquakes, the social microblog platform can be used as a better early warning system, compared to the traditional media.

5 Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the University “Ss. Cyril and Methodius”.

References

[1] Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439: 462–465. doi: 10.1038/nature04292 PMID: 16437114.

[2] Candia J, González M, Wang P, Schoenharl T, Madey G, Barabási AL. (2008) Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical 41: 224015. doi: 10.1088/1751-8113/41/22/224015.

[3] Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: User movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD ‘11, pp. 1082–1090.

[4] Kung KS, Greco K, Sobolevsky S, Ratti C (2014) Exploring universal patterns in human home-work commuting from mobile phone data. PLoS ONE 9: e96180. doi: 10.1371/journal.pone.0096180 PMID: 24933264.

[5] Sun L, Axhausen KW, Lee DH, Huang X (2013) Understanding metropolitan patterns of daily encounters. Proceedings of the National Academy of Sciences of the United States of America 110: 13774–9. doi: 10.1073/pnas.1306440110 PMID: 23918373.

[6] Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. PloS one 7: e37027. doi: 10.1371/journal.pone.0037027 PMID: 22666339.

[7] Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C. (2014) Geo-located twitter as proxy for global mobility patterns. Cartography and Geographic Information Science 41: 260–271. doi: 10.1080/15230406.2014.890072.

[8] Hui P, Mortier R, Piórkowski M, Henderson T, Crowcroft J (2010) Planet-scale human mobility measurement. In: Proceedings of the 2Nd ACM International Workshop on Hot Topics in Planet-scale Measurement. HotPlanet ‘10, pp. 1:1–1:5.

[9] Ratti C, Williams S, Frenchman D, Pulselli R (2006) Mobile landscapes: using location data from cell phones for urban analysis. Environment and Planning B: Planning and Design 33: 727–748. doi: 10.1068/b32047.

[10] Berlingerio M, Calabrese F, Lorenzo G, Nair R, Pinelli F, Sbodio ML. (2013) AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data, Berlin, Heidelberg: Springer Berlin Heidelberg, volume 8190 of Lecture Notes in Computer Science, chapter 50 pp. 663–666.

[11] Bengtsson L, Lu X, Thorson A, Garfield R, von Schreeb J (2011) Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti. PLoS Med 8: e1001083. doi: 10.1371/journal.pmed.1001083 PMID: 21918643.

[12] Lu X, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 haiti earthquake. Proceedings of the National Academy of Sciences of the United States of America 109: 11576–11581. doi: 10.1073/pnas.1203882109 PMID: 22711804.

[13] Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328: 1029–1031. doi: 10.1126/science.1186605 PMID: 20489022.

[14] Frias-Martinez V, Virseda J (2012) On the relationship between socio-economic factors and cell phone usage. In: Proceedings of the Fifth International Conference on Information and Communication Technologies and Development. ICTD ‘12, pp. 76–84.

[15] Smith C, Mashadi A, Capra L (2013) Ubiquitous sensing for mapping poverty in developing countries. In: Proceedings of the Third Conference on the Analysis of Mobile Phone Datasets. pp. 39–40.

[16] The American Red Cross, Web Users Increasingly Rely on Social Media to Seek Help in a Disaster, Press Release, Washington, DC. Aug 9, 2009. Available at http://www.redcross.org/portal/site/en/menuitem.94aae335470e233f6cf911df43181aa0/?vgnextoid=6bb5a96d0a94a210VgnVCM10000089f0870aRCRD. Accessed 11 November 2015.

[17] Earle,P, Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., and Vaughan, A. OMG Earthquake! Can Twitter improve earthquake response? Seismological Research Letters, Vol. 81, No. 2. March 2010: 246-251.

[18] Vieweg S. Microblogged contributions to the emergency arena: Discovery, interpretation and implications. In Computer Supported Collaborative Work, Feb 2010.

[19] Hughes AL, Palen L. Twitter adoption and use in mass convergence and emergency events. In ISCRAM Conference, May 2009.

[20] De Longueville-Smith RS, Luraschi G. ”OMG, from here, I can see the flames!”: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In LBSN ’09: Proceedings of the 2009 International Workshop on Location Based Social Networks, , ACM, New York, NY, USA. 2009. Pp:73-80.

[21] Lindsay B. Social media and disasters: Current uses, future options and policy considerations. Washington, DC: Congressional Research Service CRS Report for Congress, Analyst in American National Government. Sept. 6, 2011.Pp1-7.

[22] Disaster Psychiatry. Available at http://www.psychiatry.org/practice/professional-interests/disasterpsychiatry. Accessed 11 November 2015.

[23] Cooper GP, Yeager V, Burkle FM, Subbarao I. Twitter as a Potential Disaster Risk Reduction Tool. Part I: Introduction, Terminology, Research and Operational Applications. PLOS Currents Disasters. 2015 Jun 29. Edition 1. doi: 10.1371/currents.dis.a7657429d6f25f02bb5253e551015f0f.

[24] How tweets and algorithms can save lives, Aljazeera Online. http://www.aljazeera.com/indepth/opinion/2014/11/how-tweets-algorithms-can-sav-20141130142519956906.html, accessed, 05.12.2014.

[25] M. Imran, C. Castillo, J. Lucas, M. Patrick, and V. Sarah. AIDR: Artificial intelligence for disaster response. In Proc. WWW (Demos). ACM, 2014.

[26] Poplava neodgovornosti, B92 online (in Serbian). http://blog.b92.net/text/24208/Poplava-neodgovornosti/#.U3qKbuQz8nU.facebook, accessed: 01.11.2014.

[27] Leetaru, Kalev, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, & Eric Shook. "Mapping the global Twitter heartbeat: The geography of Twitter." First Monday [Online], 18.5 (2013): n. pag. Web. 5 Dec. 2014.

A Cost Effective Solution for Audio Surveillance Using Embedded Devices as Part of a Cloud Infrastructure

Elisavet Charalambous, Nectarios Efstathiou, Nikolaos Koutras
Advanced Integrated Technology Solutions & Services - ADITESS Ltd
Byzantiou 40, Nicosia, Cyprus
lc@aditess.com, nefstathiou@aditess.com, management@aditess.com

Abstract: Audio covers a 360° area day and night, at a low cost, and surpasses the limitation of the viewing of conventional surveillance cameras. The lack of audio in surveillance systems impacts significantly the ability of security personnel to act timely, if at all, in cases of emergency. In this paper we present an ethical audio surveillance system developed for the cost effective and real-time detection of auditory events of interest. The proposed system utilises a low-cost embedded system for the recording of audio and the first level of event detection, with lightweight analytics, and a virtual machine hosted on a private cloud, for the execution of second level event validation. Experimental results suggest that 8 kHz mono audio is sufficient for the real-time processing of data with not much degradation in classification performance. Classification on the embedded system is performed based on a cascaded implementation of C4.5 classifiers while audio classification on the cloud side implements four very well-known methods: C4.5, Random Forests, Support Vector Machines and k-Nearest Neighbours. The performance of the algorithms is evaluated using the classification accuracy and area under the receiver operating characteristic (ROC) curve (aka AUC).

Keywords: Audio analysis, Real-time data handling, Surveillance system, Embedded System, Cloud Infrastructure

1 Introduction

Research on automatic surveillance systems has recently received particular attention, due to the increasing importance of these systems as well as the prohibitively growing expenses as the number of deployed sensors escalates [3]. The field of audio signal classification consists of methods for extracting relevant features from a sound in order to identify into which of a set of classes the sound is most likely to fit. The feature extraction and grouping algorithms used can be quite diverse depending on the classification domain of the application [4].

In particular, the use of audio sensors in surveillance and monitoring applications has proved to be particularly useful for the detection of events like screaming and gunshots [5], [6]. Such detection systems can be efficiently used to signal to an automated system that an event has occurred and at the same time, to enable further processing like acoustic source localization for steering a video-camera. Traditional implementations involve the use of speech/music segmentation and classification [7], [8] and audio retrieval [9]. Much of the previous work about audio-based surveillance systems has concentrated on the task of detecting some particular audio events. Early research stems from the field of automatic audio classification and matching [9]. More, recently, specific works covering the detection of particular classes of events for multimedia-based surveillance have been developed. The SOLAR system [10] uses a series of boosted decision trees to classify sound events belonging to a set of predefined classes, such as screams, barks etc.

Successive works have shown that classification performance can be considerably improved if a hierarchical classification scheme composed by different levels of binary classifiers is used in place of a single-level multi-class classifier [11]. The hierarchical approach has been employed in [6] to design a specific system able to detect screams/shouts in public transport environments. A slightly different technique is used in [5] to detect gunshots in public environments. Several binary sub-classifiers for different types of firearms are run in parallel. In this way, the false rejection rate of the system is reduced by a 50% on average with respect to a single gunshot/noise classifier. Finally, in [11] a hierarchical set of cascaded Gaussian Mixture Models (GMM) is used to classify 5 different sound classes. Reported results show that the hierarchical approach yields accuracies from 70 to 80% for each class, while single level approaches reach high accuracies for one class but poor results for the others.

Despite the advances, none of the previously mentioned systems has been developed for operation on computationally and power limiting devices; imposing constraints on the complexity of deployed analytic solutions and the extraction of audio features. The proposed scheme has been deployed for the needs of the P-React project, funded by the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no 607881.

2 Problem Formulation

It is not expected an algorithm to perform equally well with any dataset. The results are highly dependent on the parameterisation of the algorithm, as well as the structure and complexity of the dataset. Since the implemented solutions aim use in open and unrestricted environments, effective operation of the systems – serving the already defined specification – require analysis of the data with multiple techniques where each one examines different aspects of the sample. Within the context of this paper, this implies the analysis of the perceived auditory scene for its voiced and unvoiced characteristics.

2.1 System Operational Information

For the needs of this study, a series of audio datasets which do not only involve events of interest but also background noise and other random sounds composing a mixture of different sound sources captured from different environments is used. This process is considered necessary as it enhances system robustness and removes any bias. Additionally, during training bootstrapping with replacement has been used for generating new datasets composed of a balanced number of samples, allowing uniform class distribution so as to avoid overfitting over a certain class.

The embedded system on which the implemented solutions run is the A10-OLinuXino-LIME equipped with an A10 1GHz Cortex-A8 ARMv7 CPU, 512MB DDR3 RAM memory [13]; a cost effective low end device which supports wired connections. On the other side, the VM hosted on cloud is allowed a CPU clocked at 2.2 GHz, 2 GB of RAM, and 20 GB of storage capacity. Both embedded and cloud systems run under Ubuntu OS while algorithm implementation is performed in C on the ES and in Python 2.7 on cloud.

Analysis of the audio stream is carried out in a block based fashion where for each block a number of spectral, cepstral and unvoiced coefficients are extracted and used for subsequent classification. For each sound block, the first 13 Mel-frequency cepstral coefficients (MFCC) [12] are retained along with the values of zero-crossings, and block energy. The performance of the system is a critical factor, therefore lightweight analytics are performed on the, power limited, embedded system. Upon occurrence of an event the results are transferred on the cloud for further analysis. This method is adopted to limit the number of false positive and therefore allowing the system to operate without unwanted traffic. The selection of the specific classification method depended on both the classification performance as well as the required processing time. The selection of the algorithm was determined based on the results of an extensive experiment involving a number of classification methods.

During operation, the ES audio analytics module performs feature extraction on discretized blocks of audio of pre-defined length. Transmission of detected events occurs in the form of metadata objects. It is expected that an audio event will raise multiple events, due to the fact that audio is processed into small blocks. The module on cloud acknowledges this and considers this as a potential manner in increasing the confidence of class prediction.

3 Methodology

There is no universal solution for every problem. The design of the methodology to tackle a problem has to serve its specific needs while considering limitations and constraints. The designation of the implemented methodology accounts for the non-uniformity in the distribution of classes, as well as the fact that audio samples significantly different from the training set may be collected during a real-life senario.

For the needs of this study, bootstrapping with replacement is used to allow the generation of new datasets; a balance between true and false events was maintained to avoid overfitting to either category. The choice of resampling adds the assumption that the dataset is representative to the population. Moreover, the performance of the algorithms is evaluated using the classification accuracy, i.e. the number of correct predictions from all predictions [1], as well as the AUC index [2]. The calculation of both indices is done in an effort to assess their robustness and emphasise the significance of their choice.

The 5x2 cross validation paired *t*-test [13] and the 5x2 cross validation *F*-test [14] were deployed to statistically test the significance of the classification results and to evaluate their robustness. Benchmarks on significance testing propose that cross validation testing methods are more robust when dealing with small datasets where reproducibility of the experiment is not an issue [13]. The 5x2 cross validation method was selected to allow large enough datasets for testing while ensuring that no further dependencies of overlapping training and testing sets are introduced when cross validation is used [15].

The methodology deployed for this paper has been introduced in [16]; variants of this scheme have been deployed both for cloud and lightweight analytics serving different objectives. On the cloud the objective is to determine which algorithm performs the best, in terms of classification evaluation metrics, while on the embedded system the objective is to find the algorithm that satisfies the trade-off between classification performance and real-time operation. In each instance, the experiment was allowed to run for 500 iterations to allow the generation of valid statistics and the significance of the results was calculated at level 0.05 ensuring that a 95% confidence level for the results of statistical testing.

4 Classification Algorithms

The C4.5 algorithm is an extension of the earlier ID3 algorithm and performs classification by generating a decision tree [17]. Decision trees are generated incrementally by breaking down a dataset into smaller and smaller subsets. C4.5 builds decision trees from a set of training data, using the concept of information entropy. At each tree node, the feature that most effectively splits the dataset into subsets is selected while the attribute with the highest normalised information gain is chosen to make the decision. The idea is to refine T (the tree) into subsets of samples that are heading towards one-member class collections of samples. An appropriate test is chosen, based on a single element that has one or more mutually exclusive outcomes [18]. The decision tree for T consists of a decision node identifying the test and one branch for each possible outcome, the C4.5 algorithm then recurs on the smaller sub-lists. The decision trees generated by C4.5 can be used for classification and it is referred as a statistical classifier.

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [19][20]. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large while it also depends on the strength of the individual trees in the forest and the correlation between them. Random forests on bagging where the combination of learning models, is thought to, increase the classification accuracy. The main idea of bagging is to average noisy unbiased models to create a model with low variance, in terms of classification; a large collection of de-correlated decision trees. Supposing that matrix S is the set of training samples, used for training the classification model, random forests generate a number of different and independent decision trees from an equal number of random data subsets (selected as random sets from the original set. Upon prediction, new uncategorised samples are classified by all trained trees. The element is assigned the label based on the majority rule of class label each tree has returned.

Support vector machine is based on statistical learning theory and the structural risk minimization principle [20]. Using the training data, SVM implicitly maps the original inputs pace into a high dimensional feature space [21]. Subsequently, in the feature space the optimal hyperplane is determined by maximizing the margins of class boundaries [22]. The training points that are closest to the optimal hyperplane are called support vectors. Once the decision surface is obtained, it can be used for classifying new data. Consider a training dataset of instance-label pairs $(\boldsymbol{x_i}, \boldsymbol{y_i})$ with $\boldsymbol{x_i} \in \Re^n, \boldsymbol{y_i} \in \{1, -1\}$and $i = 1, \ldots, m$. In the current context of audio classification, $\mathbf{x}$ is a vector of input space that contains the previously extracted audio coefficients. The two classes {1,-1} denote identified event and misidentified event. The aim of the SVM classification is to find an optimal separating hyperplane that can distinguish the two classes {1,-1} from the mentioned set of training data. For the case of linear separable data, a separating hyperplane can be defined as: $\boldsymbol{y_i}(\boldsymbol{w} \times \boldsymbol{x_i} + b) \geq 1 - \xi_i$ where $\mathbf{w}$ is a coefficient vector that determines the orientation of the hyperplane in the feature space, b is the offset of the hyperplane from the origin, xi is the positive slack variable [23].

Finally, k- Nearest Neighbour (k-NN) is a non-parametric approach, operating in the belief that a sample will more likely belong to the class of its closest already classified artefacts [24]. k-NN is among the simplest and most intuitive machine learning algorithms. For each uncategorised artefact, its Eucledian distance to all classified samples is measured, the k closest samples are selected and the artefact is categorised to the class most of the k neighbours belong; k is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its immediate nearest neighbour, according to some distance metric. The input consists of the k closest training examples in the feature space and the output is a class membership.

5 Embedded side

The audio analytics module on the embedded system (ES) is designed to emphasise on the detection of screaming, glass breaking and gun-shooting/loud explosions with the deployment of a timely efficient method. Within the tasks of the embedded system is the recording of audio from the attached microphone, its encoding in the WAVE uncompressed format, its resampling to the preferred sampling rate as well as the transmission of extracted features and lightweight analytics results on the cloud, therefore the latency caused by these aspects needs also to be considered. Additionally, due to privacy reasons, the audio data should neither be retained on the ES after a block has been processed, nor be transmitted on the cloud through the network. Therefore, only the extracted feature coefficients are transmitted to the cloud for further processing. The need of operating in real-time imposes the extra challenge in preserving low processing (feature extraction and classification) times.

Classification on the ES is performed with C4.5. The selection of this method emerged after a systematic experiment discussed in [16]; during this process the effectiveness of a number of possible audio features has also been tested. Statistical testing revealed that in most cases C4.5 was significantly better than its counterpart algorithms while in cases where the null hypothesis (i.e. algorithm A is not significantly better than algorithm B) is confirmed C4.5 was proven to be more efficient in terms of processing times.

5.1 Time Complexity

The necessity to operate and take decisions in real time requires splitting the received data stream into frames of predefined size; each frame is sequentially analysed and a set of extracted features is obtained. The prediction module is then called to make the binary decision (i.e. the sound is alarming or not). The trade-off between processing time and classification error (measured as the misclassification rate) is considered critical and as a result, the parametrisation of the algorithm is determined based on the results, shown in Table 1, involving the following parameters: audio sampling frequency, block size (expressed in ms), nfft frame, number of filter bank and MFCC coefficients, misclassification error and average (block) processing time. Based on the reported times in Table 1 the parameterisation of configuration #1 is adopted.

Parameter	*Configuration*		
	#1	#2	#3
F_S (kHz)	8	8	8
Block (ms)	140	100	140
FBank	22	22	22
MFCC	13	13	10
Accuracy	1.40%	1.60%	1.60%
Time (ms)	85.74	89.06	65.13

Table 1: Top 3 configurations obtained in an experiment which involved the parameterization of the audio analytics module.

5.2 Classification strategy

The audio analytics module on the embedded system solves the classification problem with a number of cascaded binary classification trees (shown in Figure 5); generated with the means of the C4.5 algorithm. Processing involves a number of steps depending on the output of previous steps. First the input sample goes through the three trees in Tier 1, in the case where all algorithms classify the sample as non-alerting no further processing is required, otherwise processing goes through the tree in Tier 2 which identifies between voiced and unvoiced sounds. If a scream is detected then the sample is labelled as screaming and the results along with features is queue for transmission to the cloud. In the event of an unvoiced event the algorithm moves to Tier 3 and the discrimination between the sounds of glass breaking and gunshots.

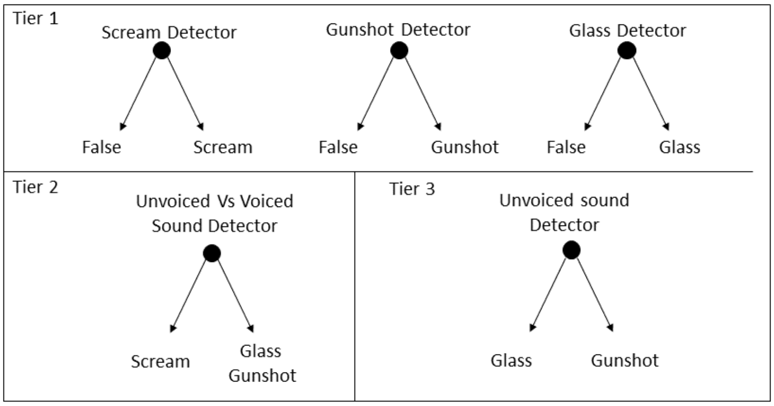


Figure 5: Lightweight analytics on ES

6 Cloud Side

On the cloud side a number of different analytics methods are available for analysis. Due to the diversity of audio sounds that may occur in trial setups, the analysis of the data from different perspectives is considered necessary, therefore the selection of the algorithms was driven by this need; each algorithm exploits best certain data attributes. Despite our selection of classification methods, the proposed design supports the addition of any other classification/clustering method.

Due to this, four different methods have been implemented: C4.5 based on decision trees, random forests, an ensemble DT method, and two statistical methods known as Support Vector Machines (hereafter SVM) and k- Nearest Neighbour (hereafter k-NN). A more complex structure, based on neural networks, known as Learning Vector Quantisation (LVQ) [26] has also been implemented, however during testing it was decided to be left out as the produced results weren’t consistent enough.

6.1 Classifier Performance Evaluation

The implemented cloud analytics algorithms for audio offer more flexibility compared to the designed strategy on the ES. The computational resources on the cloud enable the quick analysis of transferred parameters through multiple algorithms in low confidence cases.

The same approach as earlier has been followed for the evaluation of algorithms on cloud. During the experiment all algorithms operated with the same parameterisation for all generated datasets so as to avoid the introduction of dataset bias in the performance of the algorithms, however these numbers emerged as a result of several experiments. Therefore, C4.5 operated with no pruning, k-NN with k=3, random forests with tree bagging of 20 and finally SVM with the RBF kernel, a maximum value of 15000 iterations and a 5% level of violation of the KKT conditions in cases where the algorithm does not reach conversion.

For consistency purposes the evaluation of each algorithm for each type of alert involved the calculation of the following metrics: Accuracy, Sensitivity, Specificity, Precision, Recall, F-Measure, G-Mean, and AUC. In total a number of three experiments have been performed, one for each of the three possible alerts; results shown in Table 2. The experiments, tested the algorithms performance in discriminating the sounds of glass breaking, gunshots/explosions and screaming from non-scream sound (background noise, people talking, ambient sound in transportation media etc.).

	Glass Vs NonScreams				Gunshot Vs NonScreams				Screams Vs NonScreams			
	C4.5	**k-NN**	**RF**	**SVM**	**C4.5**	**k-NN**	**RF**	**SVM**	**C4.5**	**k-NN**	**RF**	**SVM**
Accuracy	1.00	**1.00**	1.00	0.81	0.99	0.98	**0.99**	0.97	1.00	1.00	**1.00**	0.99
Sensitivity	1.00	**1.00**	1.00	0.78	0.99	0.99	**0.99**	0.97	1.00	1.00	**1.00**	0.99
Specificity	0.99	**1.00**	1.00	1.00	0.86	0.93	**0.96**	0.99	0.97	0.98	**0.99**	1.00
Precision	0.99	**1.00**	1.00	1.00	0.99	1.00	**1.00**	1.00	1.00	1.00	**1.00**	1.00
Recall	1.00	**1.00**	1.00	0.78	0.99	0.99	**0.99**	0.97	1.00	1.00	**1.00**	0.99
F-Measure	1.00	**1.00**	1.00	0.87	0.99	0.99	**1.00**	0.99	1.00	1.00	**1.00**	0.99
Gmean	0.99	**1.00**	1.00	0.88	0.93	0.96	**0.97**	0.98	0.98	0.99	**0.99**	0.99
AUC	1.00	**1.00**	1.00	0.71	0.93	0.86	**0.94**	0.75	0.98	0.92	**0.99**	0.70

Table 2: Classification performance for the detection of events

7 Discussion

Results of the above experiment suggested that C4.5 was consistently the quickest algorithm both during prediction and training while it also returned high classification scores. k-NN performs best in discriminating between the sound of glass breaking and non-screaming, even though RF and C4.5 follow its performance very closely; k-NN is no significantly better than C4.5 and RF with all 3 algorithms reporting extremely high rate in both accuracy, F-measure and AUC. RF outperforms the rest algorithms in the detection of gunshots and abnormally loud sounds reporting almost perfect classification, very shortly followed by C4.5 and k-NN. However the results show that RF and C4.5 report better AUC values than k-NN. RF outperforms also in the detection of screams with very high rates, shortly followed by C4.5 and k-NN. Despite this, none of the algorithms is significantly better than the others in terms of Accuracy, Sensitivity, Specificity, Precision, Recall, F-Measure and G-Mean.

SVM constantly produces lower performance rates compared to the rest algorithms in terms of AUC. The variability between reported values of accuracy, f-measure and AUC provide a lead for further investigation of characteristics between the different classes. SVM was parametrised in a way to accept a degree of error in favour of running time. The parameterisation allows for extremely quick classification.

8 Conclusions

The fully automatic and reliable identification of sounds and alerts in real-time, at the current stage of advancement in technologies, is not possible when a computationally restricted ES is involved in the process. However, the results indicate that reliable ways of detecting alerts within an environment are feasible. Despite the high scores of the performed experiment, a confidence classification metric may be calculated as a function of the number of alerts that have occurred in the clip, the matching between ES and cloud analytics as well as the fusion of results obtained from the simultaneous analysis of the clip with multiple algorithms (possibly audio, video and/or depth).

References

[1] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[2] C. D. Brown and H. T. Davis, “Receiver operating characteristics curves and related decision measures: A tutorial,” *Chemom. Intell. Lab. Syst.*, vol. 80, no. 1, pp. 24–38, Jan. 2006.

[3] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 21–26.

[4] D. Gerhard, *Audio Signal Classification: History and Current Techniques*. 2003.

[5] C. Clavel, T. Ehrette, and G. Richard, “Events Detection for an Audio-Based Surveillance System,” in *2005 IEEE International Conference on Multimedia and Expo*, pp. 1306–1309.

[6] J.-L. Rouas, J. Louradour, and S. Ambellouis, “Audio Events Detection in Public Transport Vehicle,” in *2006 IEEE Intelligent Transportation Systems Conference*, 2006, pp. 733–738.

[7] L. Lu, H.-J. Zhang, and H. Jiang, “Content analysis for audio classification and segmentation,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 504–516, Oct. 2002.

[8] PINQUIER Julien, ROUAS Jean-Luc, and ANDŔE-OBRECHT Ŕegine, “Robust Speech/Music Classification in Audio Documents.,” in *7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2002.

[9] T. Zhang and C.-C. J. Kuo, “Hierarchical system for content-based audio classification and retrieval,” in *Photonics East (ISAM, VVDC, IEMB)*, 1998, pp. 398–409.

[10] D. Hoiem and R. Sukthankar, “SOLAR: Sound Object Localization and Retrieval in Complex Audio Environments,” in *Proceedings. (ICASSP ’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 5, pp. 429–432.

[11] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli, “Audio Based Event Detection for Multimedia Surveillance,” in *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, vol. 5, pp. V–813–V–816.

[12] L. R. Rabiner and R. W. Schafer, “Introduction to Digital Speech Processing,” *Found. Trends® Signal Process.*, vol. 1, no. 1–2, pp. 1–194, 2007.

[13] T. Dietterich, “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms.,” *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Sep. 1998.

[14] E. Alpayd, “Combined 5 × 2 cv F Test for Comparing Supervised,” *Neural Comput.*, vol. 1892, no. 8, pp. 1885–1892, 1999.

[15] S. Salzberg, “On Comparing Classifiers : Pitfalls to Avoid and a Recommended Approach,” *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 317–328, 1997.

[16] E. Charalambous, M. Dikomitou-Eliadou, G. M. Milis, G. Mitsis, and D. G. Eliades, “An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relationships,” *J. Archaeol. Sci. Reports*, Aug. 2015.

[17] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer New York, 2013.

[18] B. Thakur and M. Mann, *Data Mining with Big Data using C4.5 and Bayesian classifier. International Journal of Advanced Research in Computer Science and Software Engineering*. 2014.

[19] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[20] Vladimir N. Vapnik, *Statistical Learning Theory*. 1998.

[21] M. Kanevski, A. Pozdnoukhov, and V. Timonin, *Machine Learning for Spatial Environmental Data: Theory, Applications, and Software*. 2009.

[22] S. Abe, *Support Vector Machines for Pattern Classification*. London: Springer London, 2010.

[23] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, vol. 9. John Wiley & Sons, 2012.

[25] S. L. Salzberg, “C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993,” *Mach. Learn.*, vol. 16, no. 3, pp. 235–240, Sep. 1994.

[26] T. Maps and K. Networks, “Self Organizing Maps : Fundamentals What is a Self Organizing Map ?,” pp. 1–15, 2004.

Semantic Technologies for the Masses

Georgi Pavlov, Pavel Genevski

SAP Labs Bulgaria, EOOD

Blvd. Tzar Boris III 136A, Sofia, Bulgaria

georgi.pavlov@sap.com; pavel.genevski@sap.com

Abstract: Semantic technologies have been a focus area in the research field for long and have received significant attention particularly in the last decade. While there is a general agreement that they can play important role in driving the software world to a whole new level, they still remain largely a niche technology. This article discusses the drivers to a more massive adoption by end-users and technology providers, the challenges that currently hinder this progress and some lessons learned from applying semantic technologies in practice.

Keywords: Semantic technologies, adoption metrics, databases, in-memory computing, services.

1 Introduction

Semantic technologies is a term used to collectively refer to the technologies for building and using the Semantic Web [1], which is the W3C vision for a "Web of (linked) Data", enabling machines to make sense and use of the enormous amounts of data on the Web and boost its value to unprecedented levels. The Semantic Web is an evolution of the World Wide Web (Web of Documents) and builds upon the very same principles, using Internet as underlying global integration technology. However, unlike the blooming of the Web of Documents in its infant years, we never faced similar explosive adoption in the Web of Data domain. In the next sections we shall discuss the challenges in the mass adoption of the technology and how that affects its progress and future.

2 The Semantic Web - past and present

Looking at the Semantic Web history, there are several milestone moments that standout. The first moment the **Semantic Web inception**. Technologies employing semantics for building knowledge systems have been around even before computers were invented. But the idea to exploit them on the Web has been born right about after the invention of the World Wide Web in the last century 90s. Formally, the term and original vision for the Semantic Web and its key technologies are first mentioned in 2001 in an article co-authored by sir Tim

Berners Lee [2]. The Semantic Web was not a self-contained concept, but followed a consistent approach, leveraging existing Web standards, such as XML, XMLSchema and RDFSchema, to integrate into the W3C standardization efforts in establishing the future of the Web. In the next few years more and more technologies (OWL, SPARQL) were added to satisfy different aspects of the vision for Web of Data that was getting more and more ambitious in scope too. The good thing about this period was that the Semantic Web was a first-class citizen in the vision for the Web. The hype around it was inflating large and fast, fueled by speculations about the possibilities it could open. The negative was that it was part of a vision that had hard times making it to adoption. It is not the first time when reality mismatches with the standardization bodies vision. Wellknown examples are REST APIs versus WebServices, JSON versus XML and many others. The obvious common pattern is that practical aspects, such as simplicity, are often neglected, but technologies built with them in mind are much more potent for adoption. Following these concerns, around 2006 Tim Berners Lee started to ‘rebrand’ the Semantic Web to feel more native into the realworld Web where it was supposed to ‘live’ not as an overlay but as integral part. This next major milestone was **the raise of the Linked Data**. It came around about the time when the lack of real and massive adoption of the Semantic Web term, in contrast to its claimed potential, was questioning the feasibility of the idea. Effectively the data was still not on the Web. And neither did the technology face much of an industrial adoption, beyond a few domains making it mostly a risky niche. The Linked Data was first mentioned around 2006 by Tim Berners Lee [3] but its more influential establishment was made by another article [4] that he coauthored in 2009 and his emotional and inspiring TED talk [5] on the topic. This was a very sensible move that made the idea relevant and opened up real adoption opportunities. As a consequence a real economy of IT for the Open Data appeared and started to gain momentum pioneering in UK and USA in the public sector domain. Around that time we witnessed also the first large industrial adoptions. But it was not a miracle that happened over the night. The key success was that this transition managed to build real life examples that exploit the benefits of the technology. This shows an important prerequisite for the mass adoption of a technology, namely its relevance. The move to Linked Data was a huge step forward but not the last one. It revealed a number of challenges that wait for their solution before the technology can become a widely adopted ‘commodity’ technology, such as conventional programming for the Web.

The W3C organization and vision for the Web has also changed since the early days. Currently, it reflects much better the reality, by shifting from imposing standards and practices that are not yet adopted to those that are already widely employed, such as REST and JSON for example. Building upon adopted, popular standards was indeed a smart and very necessary move for the survival of the Semantic Web technologies.

The W3C activities around the Semantic Web have been reformed into a new Data Activity [6] working in to major directions. One is technology, focusing on the key pillars Linked Data, Vocabularies, Query and Reasoning. The other direction is Vertical applications.

Considering the rich history of ups and downs so far, there are lots of lessons that can be learned for the technology progress and adoption. We can also outline parallels to other technologies and using the pattern to reason on the right approach at successful massive adoption.

3 Reaching out to the masses

There are two major goals for a new technology that seeks establishment. Goal number one is to survive, winning a critical mass of adopters to selfsustain. Goal number two is to reach market saturation, i.e. maximum adoption. It is necessary to understand where the Semantic Web technologies are in this respect, what slows them, and what can drive them further to better adoption, using a methodology.

First of all, a technology provider needs to be wellaware of the group that will adopt the new technology. In fact, an adopting group is not homogenous but consists of several subgroups, each with its own interests and therefore requires distinct strategies at approaching them. Winning maximum individuals in these subgroups directly affects the ability to reach the market share saturation goal. The theory of the dynamics of the process of induction of innovations in a target group is studied in the popular research Diffusion of Innovations [7], which identifies five groups with (mostly) different share.

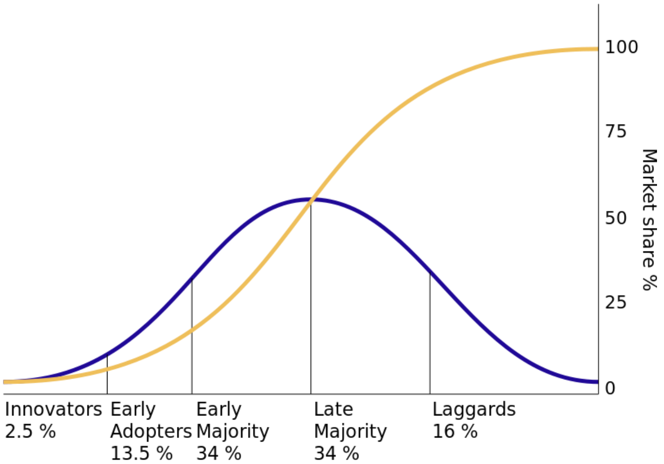


Figure 1: The diffusion of innovations **(source: https://en.wikipedia.org)**

Despite the diverse driving factors, a common aspect among all groups is the maturity of the technology and it has an effect with different intensity. One way to look at maturity is the Gartner hype curve, which provides a graphic representation of the maturity and adoption of technologies and applications [8]. It employs maturity vs visibility as axes but this can be seen also as expectations vs time. With all the conditionals around how punctual is the method, it shows a commonly accepted overview of the maturity, adoption and social effect of a technology. A more precise method is the Technology Readiness Level assessment [9]. But it does not reflect the social interest and adoption. In [10], Gartner explains the hype indicators that form the curve (Figure 2) and also provides its own maturity assessment scale with indications for the expected market penetration in different phases.

The union of the methodology for identifying adoption groups and the hype curve can tell us, which is the most concerned group on each stage of maturity level and interest. Based on that forming the right message and approach is more reliable.

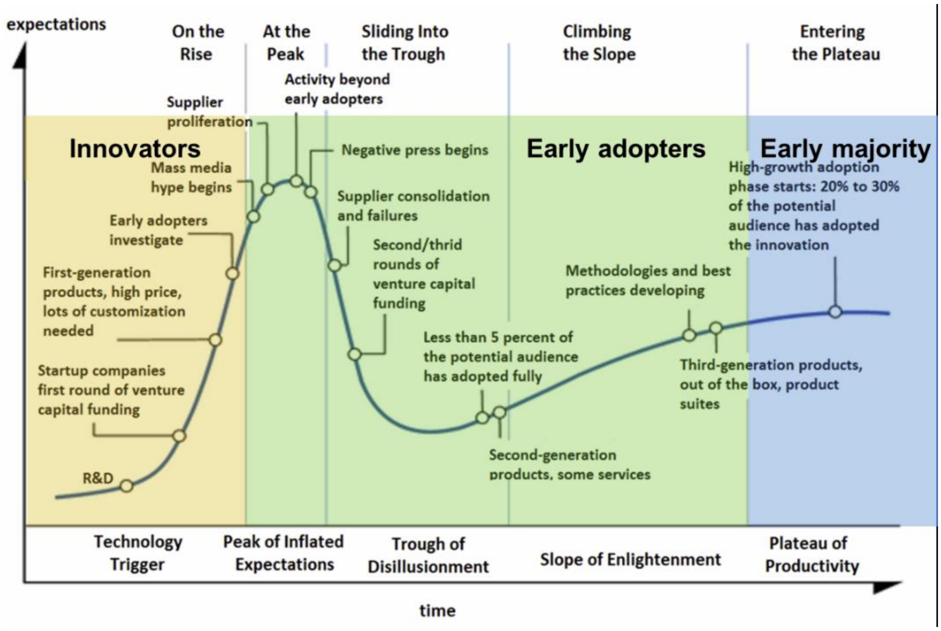

expectations
On the Rise
At the Peak
Sliding Into the Trough
Climbing the Slope
Entering the Plateau
Activity beyond early adopters
Supplier proliferation
Innovators
Early adopters
Early majority
Negative press begins
Mass media hype begins
High-growth adoption phase starts: 20% to 30% of the potential audience has adopted the innovation
Supplier consolidation and failures
Early adopters investigate
Second/thrid rounds of venture capital funding
Methodologies and best practices developing
First-generation products, high price, lots of customization needed
Less than 5 percent of the potential audience has adopted fully
Third-generation products, out of the box, product suites
Startup companies first round of venture capital funding
Second-generation products, some services
R&D
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

Figure 2: Adopter types on Gartner's hype cycle indicators curve

Unfortunately, not all groups fit in the union as evident on Figure 2. Still we can conclude the following. The relatively small (2.5%) group of enthusiastic innovators is the one caught up in the hype of a new trendy technology despite its immaturity. The selling point for this group is technology coolness, not technology readiness. As speculations drive the interest towards its peak a larger group get involved. The one of the early adopters (13.5%), which normally would be visionaries that are among the first to open new directions and introduce major differentiators. They carry out the hard work of living up through the peak of inflated expectations, then the lows of disappointment, and if the technology makes it to there, then through slow uptake on the slope of enlightenment to realize what actually you can do with this technology in a measurable way. They can live with the risks and the process of maturing should there be enough potential to make a difference. Once this is clear, the first major group steps in. The group of early majority (34.5%) is all about pragmatism and adopts the novel approaches from visionaries into industry as long as risks can be mitigated sufficiently and the technology has proven, measurable value.

According to Gartner hype curve for the Web domain, the Semantic Web currently is close to leaving the Through of Disillusionment phase. Indeed, there is relatively small adoption and a second generation of products started to appear with first attempts for cloud services, such as linkeddata.center and dydra. The hype that would attract enthusiastic innovators is long gone taking with it the approach from the early days of trying to fit the technology into a vision disconnected from reality. After a bit of criticism and grounding of expectations the course was refactored towards Linked Data to make it a real Web citizen without major disruption by simply using what is available even today.

So what can we conclude about the maturity and target adopters as of today? Now it is time for a sober rethinking of how the technology can become relevant in the real world and mature. Gartner estimates another 5-10 years period for that to eventually happen, which is among the longest periods. Partly, this is probably due to its tack of record for relatively slow movement towards maturity so far. On the other hand the technology itself is way more complex and thus not comparable to e.g. WebServices and it is natural to expect longer timelines. Early adopters with visionary profile are expected to step in and open up compelling opportunities for solutions based on semantic technologies, bringing it closer to the expectations of real world adopters. One consequence could be that instead of trying to create a global network we are first going to see successful and convincing experiments on smaller scale within organizations in particular domains. The globalization and networking may eventually come on a later stage if and when feasible. Or never. The next section discusses possible application areas that have proven successful to experiment with semantics so far in that direction.

4 Reaching out to the masses

Semantic or technologies sharing similar concepts have already been successfully deployed in a variety of industry sectors to solve high value problems or reach out to large customer base. This section provides an overview of a few examples.

4.1 Intelligence

During the WWII, the US FBIS (later part of CIA) was responsible for surveying foreign broadcasts and extracting the meaning and intent out of them into “spot bulletins” [11]. The analysts had to collect and organize facts about the activities and relationships between given entities (individuals and organizations) [12]. With the advent of computers and application of semantic nets, this work was automated to some extent as evident form the portfolio of the defenserelated SDC Corporation [13]. The publicly available information on how exactly semantic nets are used in the intelligence apparatus is scarce but many sources indicate undeniably the use of knowledge graphs in this community. For example, the “Digital Reasoning” company, that is part of the In-Q-Tel portfolio is concerned with the extraction of structured information from human communication, such as emails and storing it in a “private Knowledge Graph” [14]. Another report reveals many technical details about the scale of problems that the graph DBMS systems at NSA have to deal with [15].

Semantic technology is also slowly making its way into the commercial business intelligence domain as well, where it is used for OSINT and monitoring of competitors [16] [17]. Natural language processing, speech recognition, image recognition, face detection etc. are often used in order to efficiently transform the raw data feeds (e.g. HTML documents, voice, pictures) into structured information, from which facts can be extracted into a knowledge graph.

4.2 Internet giants

Google’s mission statement “organize the world's information and make it universally accessible and useful” [18] suggests that the company’s search engine must understand the semantics of the information in order to deliver on this promise. Early attempts in that direction was the technique of “Latent semantic indexing” [19], which they put into production in 2004. The next notable addition was making search results dependent on the user’s search term history, understanding the user interests to search hits more relevant. The addition of “Universal search” followed in 2007 and the search engine started to distinguish between different web resources such as Video, Images, News, etc. The biggest publicly known steps that Google made into the direction of semantic technologies, however were the schema.org [20] effort and the following acquisition of the Metaweb technologies and its Freebase product, which became the Google Knowledge graph [21]. In 2014 the Knowledge Graph was renamed to Knowledge Vault, with the main difference stated that the new technology is using machine learning to extract facts from unstructured data and therefore uses a probabilistic element in order to denote the probability of a fact being true.

Microsoft’s Bing search engine employs a similar technology called Satori. It is a knowledge graph that collects information about different entities and the relationships between them. Satori populates the knowledge graph with information that people search most often, such as celebrities, colleagues, popular places etc. using LinkedIn, Wikipedia, Facebook, Twitter, Klout etc., and unifies the semantics behind these heterogeneous data sources.

Facebook also employs a technology similar to other companies’ knowledge graphs, called Unicorn [22]. Facebook extracts information about the relationship between entities, observing the users’ behavior. A case from 2011 [23] shows that Facebook attempts to extract structured information by making inferences from every data point that they get access to, correlating data from different sources such as timestamp and location [23]. Advances in image recognition, speech recognition and natural language processing will tremendously increase the volume of structured information available for machine processing and will speed up the adoption of semantic technologies in this domain even further.

4.3 Bioinformatics

The need to organize information have grown up significantly in the field of bioinformatics, sometimes also referred to as synthetic biology, molecular biology etc. Recent advances in DNA sequencing technology have led to increasing data volumes and complexities in the field. The need to organize this information has been reflected in numerous efforts, a more recent of which is the SBOL standardization effort [24]. Its first goal was to come up with an interchange format, describing the chemical structure of DNA, RNA and protein building blocks. Version 2.0 of the standard extended the scope to functional aspects as well, i.e. how the building blocks may interact with each other. The SBOL standard is based on existing semantic technology components, such as URIs and ontologies. An example for a database utilizing SBOL is the SBPkb [25].

5 Enterprise concerns and alternatives

What are the missing pieces that keep adopters away from the semantic technologies? First of all, unless they don’t solve billion dollar problems, no one wants an investment that disrupts the existing IT infrastructure, which probably has already heavy investments. How to fit in is a big concern and topic. Consider for example the topic for integration with a Single Sign-On system that is probably already in place. Or how to comply with the company policy on protection and privacy. There is some research, e.g. [26] and [27] and some support is implemented in a few advanced RDF storage systems (e.g. OpenLink Virtuoso), but this is by no means trivial to apply yet, even when there is such option. There is a number of other enterprise concerns, some of which captured in [28]. We would like to add to that list some more concerns that we consider no less critical for enterprise adoption (very simplified) such as legal (licenses), sustainability, provenance and trust of vocabularies, change management, price (cost/value ratio), shared knowledge availability in community best practices, etc. These do not concern only the semanticenabling software but also the vocabularies as a key pillar.

Vocabularies are foundational to semantic applications development. Selecting the right ones is a mission critical task. Unfortunately, it is not easy to find out and asses their enterprise qualities mentioned above. The gap between this insight demand and supply is particularly visible if we compared indexing engines such as Openhub and Datahub. There is also the data reconciliation or ‘cleansing’ of data problem. With time it is expected that the seriously supported datasets will surface and outlive others and there will be fewer conflicts as there will be few major de facto standard ontologies per domain. Some research efforts such as [29] and [30] described possible approaches at assisting data engineers in picking the most popular vocabularies in order to build on the ‘wisdom of the crowd’ as a natural process. Large initiatives such as those delivering schema.org as commonly agreed vocabulary among the major Web search engines, or Goodrelations that is a de facto standard for describing business entities, or Linked USDL for faceted description of business services are popular examples of becoming the de facto standard in a domain and interlinking without conflicts with other standards. There is a good prospective that these vocabularies will sustain and continue to develop and draw users because they are actually used in real business.

Overall, the semantic developer does not enjoy a toolbox comparable to the one of a conventional application programmer. The entry barrier is still quite high, the choice of tools compliant with the outlined concerns - limited. That may start changing with the growing popularity of **Graph Databases.** They are reported to be at the Peak of the hype cycle curve for information infrastructure for 2015 by Gartner [31]. What drives this interest is the trend for analytics with BigData that generates a necessity for efficient and fast insight in huge data volumes that graph technologies happen to be very suitable for. From a certain perspective this can be seen as a focused application of semantic technologies into the analytics domain. We can actually observe that even some longtime vendors were quick to rebrand their offer focusing on the Graph capabilities that under the hood happen to be RDF and SPARQL. This is an excellent example of finding a reasonable and winning application of the semantic technology with great prospective for adoption. Indeed, focusing first on the very tangible problem that most enterprises have today, to capitalize on the huge amounts of data they own and have absolutely no idea about what treasures they can hide, and only then expand on the benefits of standardized technology and utilization of global knowledge is the strategy that has much more chances for success than the other way round. Of course, not all graph databases have standard semantic technology capabilities. Not at all. But that is not the main concern on this stage. In fact, they may present a very valid alternative to the standards and become defacto standards and in shorter time due to all lessons learned. We have seen a similar phenomenon with WebServices and REST APIs before. We already know that what will follow next for graph databases and it is a bit of decline in interest. But being a very potent technology solving a very relevant and critical problem, it will most likely live through this period fast and face some standardization and quite some adoption. That said, semantics can come in a much unexpected ways to the masses – through databases that have not much in common with the standards except the graph concept. Efficiency over standards.

6 Conclusions

Semantic technologies are the W3C standardized technology base supporting the vision for a global Web of data. The hype peak is behind them and they are slowly progressing towards maturity and with that to a larger consumer base too. There are several distinguished drivers that will eventually move them to the desired stage in which the major consumer groups of Early Adopters and especially Early and Late Majority will take over.

One such driver is focused semantic solutions solving business critical problems of high value. Examples are realizations in the domain of Web search engines, bioinformatics, intelligence and others. Or Big Data analytics that is currently at the peak of its hype. Some of these examples also show another trend and possibility. Semantic approaches may make it to wide adoption not necessarily be means of the standardized by W3C technologies for that purpose. For example general graph database management systems are a different technology that may serve the same purpose. It is hard to predict the future of both approaches. Analogies form history of other technologies that appeared similarly, such as WebServices and REST APIs, show that at the end both have use cases and predictions that only one will survive are still not true after a decade.

Another major driver is sufficient enterprise readiness of a critical mass of semantic technology products. The semantic developer toolbox must become comparable to the one that ‘normal’ fullstack app developers enjoy, convincible enough to jump on it for the next project. There must be clear paths for technology convergence with latest and greatest in inmemory databases, cloud computing, BigData and contemporary web programming. Finally, it is necessary to witness a vibrant and large community sharing examples to follow and learn from the wisdom of the crowd.

Gartner predicts another 5-10 years for the Semantic Web to reach productive maturity. But we have to be ready and accept that it may not be built with the same technology stack that was originally designed for that. Or at least that we shall move to that stage incrementally with not so global semantic solutions and more conventional technologies at the beginning. At the end, it is not the technology, but the solution value that should drive choices.

References

[1] http://www.w3.org/standards/semanticweb/

[2] Berners-Lee T., et.al., (2001). The Semantic Web, *Scientific American*, May 2001, 29-37.

[3] http://www.w3.org/DesignIssues/LinkedData.html

[4] http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

[5] https://www.ted.com/talks/tim_berners_lee_on_the_next_web

[6] http://www.w3.org/2013/data/

[7] Rogers E. (2003). *Diffusion of Innovation* (5th edition). Simon and Schuster

[8] http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

[9] https://en.wikipedia.org/wiki/Technology_readiness_level

[10] Fenn J., Raskino M. (2011). *Understanding Gartner's Hype Cycles,* http://www.donkersct.nl/downloadables/understanding_gartners_hype.pdf

[11] http://www.foia.cia.gov/sites/default/files/FBIS_history_part1.pdf

[12] http://www.hourofthetime.com/1-LF/November2012/Hour_Of_The_Time_11012013-The_Invisible-Government-David_Wise.pdf

[13] https://en.wikipedia.org/wiki/Semantic_network

[14] https://www.iqt.org/portfolio/

[15] http://www.pdl.cmu.edu/SDI/2013/slides/big_graph_nsa_rd_2013_56002v1.pdf

[16] http://www.expertsystem.com/solutions/defence-intelligence/

[17] http://www.old.anbr.ru/view_publication.php?lang=1&id=4

[18] https://www.google.com/about/company/

[19] http://blog.hubspot.com/marketing/google-algorithm-visual-history-infographic

[20] https://en.wikipedia.org/wiki/Schema.org

[21] https://en.wikipedia.org/wiki/Knowledge_Graph

[22] https://research.facebook.com/publications/371577712972872/unicorn-a-system-for-searching-the-social-graph/

[23] https://news.ycombinator.com/item?id=3377018

[24] http://sbolstandard.org/

[25] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3044748/

[26] Pavlov, G. (2009). *Access Control on Shared Ontologies*, IBIS 9 , 77-89

[27] Pavlov, G. (2009). *Fine-Grained Access Control on Shared Knowledge,* International Conference AUTOMATICS AND INFORMATICS '09, John Atanasoff Society of Automatics and Informatics (SAI), Sofia, Bulgaria, October 2009

[28] http://ontotext.com/documents/white_papers/The-Truth-About-Triplestores.pdf

[29] http://stasis.sunderland.ac.uk/

[30] Pavlov, G. (2009), *Repository Infrastructure for community-driven ontologies management support*, In Proceedings of First International Conference on Software, Services & Semantic Technologies (S3T), Sofia, Bulgaria, October 2009, pp. 28-29, ISBN 978-954-9526-62-2

[31] https://www.gartner.com/doc/3112217/hype-cycle-information-infrastructure-

An Innovative Digital Forensic Tool Assisting Evidence Analysis in Cyprus

Elisavet Charalambous, Romaios Bratskas, George Karkas,
Andreas Anastasiades, Nikolaos Koutras
Advanced Integrated Technology Solutions & Services - ADITESS Ltd,
Byzantiou 40, Nicosia, Cyprus[1]
Cyprus Police – Criminal Investigation department - Special cyber-crime unit[2],
Police Headquarters. Antistratigou Evagelou Floraki Str., Nicosia. Cyprus
lc@aditess.com[1], rb@aditess.com[1], gkarka@police.gov.cy[2],
aanastasiades@police.gov.cy[2], management@aditess.com[1]

Abstract: Digital forensic examination is proven to be a time consuming task that relies heavily on the qualities and extensive expertise of the Digital Forensic Examiner (DFE). Nowadays, due to the rapid increase on the volume of digital evidence, the skills of the DFE are not sufficient. As a result, multiple Digital Forensic Tools (DFTs) have been developed around the world by a number of companies aiming to assist the DFE and accelerate the examination process or interpret the data examined. The DFE has to select which DFT is more appropriate according to its functionality, limitations, complexity and cost. Authorities in Cyprus deploy a range of advanced DFTs in order to cover their needs in large scale investigations. In this paper, a new DFT developed by ADITESS for the automatic extraction of embedded information present in the form of EXIF metadata in images is presented with the aim to provide functionality not available in other DFTs. The tool has been designed based on end-user requirements and assists the examination process with the provision of temporal, spatial and contextual information.

Keywords: Forensics Examination, Big Data handling, Decision, Data fusion

1 Introduction

Digital forensics is defined in [1] as: “the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitation or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations”. Forensic examination often involves examination of large amounts of data where in most cases the examiner is being called to reconstruct snapshots of the scene in the effort to assist the inference of conclusions. Moreover, the substantial growth in the use of digital data and smartphones, imposes challenges on heterogeneous big data handling, while also enables advancements in the field.

Computer crime investigators and digital forensic tool developers, seek solutions to real-world problems [2]. Accurate scene reconstruction is rarely possible, each piece of available data only captures a short snapshot/instant of a certain aspect, potentially, related to the scene. For the effective analysis of digital forensic evidence, the examiner needs to combine different types of information admitting not only contextual but also spatial and temporal information.

Investigators usually do not lack data, but often struggle with transforming the data into investigative knowledge. A key factor of the digital investigation process is the mapping of incidents from different sources on a single platform with the aim to transform evidence in the appropriate form for use in secondary investigation aspects [3]. Moreover, transforming data in a convenient form for analysis is frequently time consuming and error prone, while performance degradation is noted over time as the software is called to cope with continuously increasing amounts of data.

This idea is illustrated in Figure 6**Error! Reference source not found.**, where data is perceived as symbols, information as the outcome of data being processed and knowledge as the product of interpreted information. The use of knowledge in the inference of conclusions and explanations forms what is referred to as understanding, while evaluated understanding leads to wisdom; the ability to perceive and evaluate long run consequences of behavior.

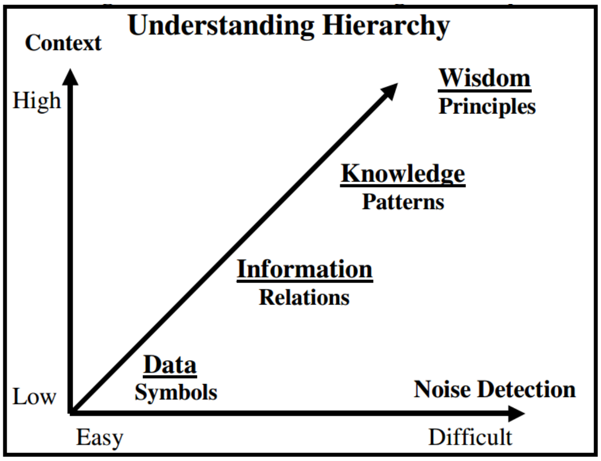


Figure 6: KMS Understanding Hierarchy [4]

Despite advances in information systems, DFEs still need to overcome a steep learning curve with respect to the core body of knowledge along with the skills of conclusion inference. Experience serves as a mediating construct, such that more experienced or expert individuals will more easily acquire awareness, discern relations, patterns, and principles, grasp the appropriate context, and detect and ignore noise in favor of meaningful inputs [4].

2 Data for Digital Forensic Analysis

Data for digital forensic analysis comes in different forms with largely varying amounts of information. The reliable extraction of conclusions requires cross-validation of information usually gathered from heterogeneous data sources. Additionally, modern digital crime scenes frequently involve multi-terabyte data stores, mission critical systems (with no option for offline imaging), ubiquitous sources of volatile data, as well as enterprise-level complex incidents in which the scope and location of the evidence are difficult to ascertain [5]. It is therefore safe to say that digital forensic investigators seek to understand past human behavior in the digital realm by studying digital artifacts.

Common sources of digital artifacts vary from multimedia content (audio, video, images) to text, cached files, system logs as well as information hidden inside cookies for which their effective analysis also needs the consideration of metadata. In this paper, we are concerned with metadata information enclosed in the encoded image files and with EXIF data in particular. EXIF is known as the standard that specifies the formats for images, sound, and ancillary tags used by digital cameras (including smartphones), scanners and other systems handling image and sound files recorded by digital cameras [6]. The analysis of image metadata may provide useful temporal and spatial information hidden behind fields admitting the device's location, time of capture, environmental conditions as well as device properties and settings; a list of captured metadata is shown in Table 3.

Table 3: Data encapsulated in images in the form of EXIF data

EXIF metadata attributes		
Filename	Date Time Digitized	GPS Latitude Ref
Make	Components Configuration	GPS Latitude
Model	Exposure Bias Value	GPS Longitude Ref
Orientation	Metering Mode	GPS Longitude
X-Resolution	Flash	GPS Altitude Ref
Y-Resolution	Focal Length	GPS Altitude
Resolution Unit	Flashpix Version	GPS Timestamp
YcbCr Positioning	Color Space	GPS Datestamp
EXIF IFD Pointer	Pixel X Dimension	Thumbnail Compression
GPS IFD Pointer	Pixel Y Dimension	Thumbnail Orientation
Exposure Time	Interoperability IFD Pointer	Thumbnail X-Resolution
F-number	White Balance	Thumbnail Y-Resolution
ISO Speed Ratings	Digital Zoom Ratio	Thumbnail Resolution Unit
Exif Version	Interoperability Index	Thumbnail JPEG Interchange Format
Date Time Original	Interoperability Version	Thumbnail JPEG Interchange Format Length

3 Digital Forensic Analysis through a Test Case

DFTs developed over the years have as their main objective to assist experts during some stage of a forensic investigation, however, they are implemented from the IT perspective rather than the law enforcement standpoint. Forensic tools can be discriminated depending on the platform they target use, EnCase, Forensic Toolkit (FTK), X-Rays forensics, NUIX, Internet Evidence Finder (IEF) and X1 Social Discovery target digital evidence examination on PC based systems, on the other hand, Cellebrite Touch, XRY, Oxygen and Mobile Phone Examiner Plus (MPE+) are designed mostly as standalone portable devices for mobile phones.

Despite the fact that such solutions have been designed for use in most standard forensic examinations, they differ mostly in functionality and, as a consequence, usability. Therefore, the available solutions adhere different levels of complexity, cost and presentations for data reporting.

EnCase Forensics v7 acquires and examines data quickly from computers, smartphones and tablets. The software is capable of increasing confidence in the reported results by complying with proven court-referenced digital forensics standards. EnCase is also capable of uncovering more potential evidence through its advanced search capabilities, boosting productivity by previewing results as data is being acquired while the DFE may search and analyze multiple evidence simultaneously. EnCase Forensics also offers a scripting solution known as EnScript which allows the automation of common investigative tasks by the DFE as long as they have the appropriate skills.

AccessData Forensic Toolkit (FTK) allows DFEs to handle massive datasets and is widely known for its ability in performing advanced code breaking and password recovery, granting access to protected data, as well as supporting functionalities like powerful searching, generation of supplemental registry reports and remote machine analysis. The built-in data visualization and explicit image detection technology allows quickly discerning and reporting the most relevant material to the analyst; also supports the generation of timelines indicating the sequence of events as these are recorded in the available artifacts. Moreover, FTK may automatically identify pornographic images through its Explicit Image Detection (EID) functionality, and apart from recognizing flesh tones, it also recognizes patterns and shapes that seem to be relevant to pornographic content while it also supports Microsoft's PhotoDNA [7]. The Access Data Forensic Toolkit (FTK) is recognized around the World as the Standard Digital Forensic Investigation Solution and it is the predominantly used tool by many forces around the world.

X-Ways Forensics is a lightweight fully portable tool operating under Windows without the need of installation. X-Ways Forensics is not resource-hungry and it is capable of retrieving deleted files and search relate information.

NUIX searches and analyzes data across multiple sources and large volumes with unmatched speed. NUIX automatically pre-filters evidence to focus on the most prospective sources offering powerful search and analysis capabilities in the hands of case investigators.

In contrast to other tools, the Internet Evidence Finder (IEF) is designed to work with computer forensics tools like EnCase, FTK, Nuix, X-ways as well as mobile forensics tools like Cellebrite UFED, XRY. As a result, examiners use IEF in combination with these tools to find more digital evidence. Finally, X1 Social Discovery is designed to extract data from the leading social media networking sites (Instagram, Facebook, Twitter and LinkedIn) as it is capable of crawling, capturing and instantly searching content from websites, webmail and YouTube.

Moving on to solutions for mobile platforms, UFED Touch by Cellerbrite is a comprehensive, standalone mobile forensic extraction device that combines outstanding mobile device support with unrivaled data extraction technology. With its intuitive GUI and easy-to-use touch screen, the UFED Touch enables physical, file system, and logical extractions of all data and passwords, included deleted data, from the widest range of mobile devices. The UFED InField Kiosk features an interface which makes extracting live device data simple while the UFED Link Analysis unifies, correlates and analyzes large volumes of different data from multiple data sources on a single platform.

Another tool called XRY allows DFEs to perform secure forensic extraction of data from a wide variety of mobile devices. The XRY Kiosk has been designed for first responders and allows quick and easy data recovery from mobile devices while the XAMN interface allows users to view and compare the contents of different devices simultaneously in one place.

Oxygen Forensic Suite Enterprise can extract device information, contacts, calendar events, SMS messages, event logs, and files. Mobile Phone Examiner Plus (MPE+) provides enhanced smart device acquisition and analysis capabilities and allows data extraction from iOS and Android devices extremely fast.

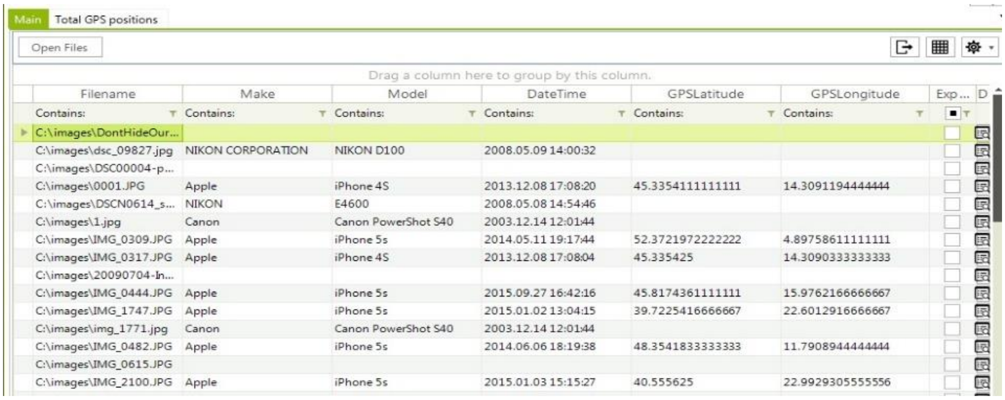
Despite the advanced features and analysis functionalities of already developed tools, it is often seen that they lack the capability of combining the different bits and pieces of a collection of images so as to allow the reconstruction of the scene. Tools striving in minimizing noise and maximizing contextual information are needed to allow the transformation of information to investigative knowledge.

4 The EXIF analysis tool

Currently, no, widely known, DFT has been designed for the utilization of EXIF information stored in images; existing commercial DFTs only display EXIF details of individual files. As a result, it is not easy for an examiner to correlate common EXIF data between images and gain spatial context; despite GPS coordinates being recorded. The DFE is therefore required to utilize scripting languages, if skilled, to be able and custom gain this insight from the data; a time consuming and error prone procedure. Due to the complexity of this procedure, often DFEs opt to display, important to the case, information after individually analyzing each artifact for the production of the forensic report, discarding, at this stage, information that might potentially be proven useful at a later stage in the investigation.

Having identified these gaps, we have developed the "EXIF Analysis" tool aiming to assist experts by allowing the efficient visualization, categorization and filtering of artifacts in large scale investigations. EXIF analysis produces intuitive representations, full customization in the order and displaying status of the various fields while still preserving all available information for further investigation.

EXIF details are seamlessly extracted from images and, by default, displayed in the form of a table where each row depicts information from individual artifacts and each column metadata attributes (see Figure 7 and **Error! Reference source not found.**). The tool offers flexibility, allowing the user to select and reorder the visible fields at each instant while perspective settings (ordering of columns and operating preferences) are stored for subsequent uses of the tool. This allows each DFE to design the perspective they find most suitable, not having to re-adjust parameters upon re-initialization of the application.



Filename	Make	Model	DateTime	GPSLatitude	GPSLongitude	Exp ...	D
Contains:	Contains:	Contains:	Contains:	Contains:	Contains:	■	
C:\images\DontHideOur...						☐	
C:\images\dsc_09827.jpg	NIKON CORPORATION	NIKON D100	2008.05.09 14:00:32			☐	
C:\images\DSC00004-p...						☐	
C:\images\0001.JPG	Apple	iPhone 4S	2013.12.08 17:08:20	45.3354111111111	14.3091194444444	☐	
C:\images\DSCN0614_s...	NIKON	E4600	2008.05.08 14:54:46			☐	
C:\images\1.jpg	Canon	Canon PowerShot S40	2003.12.14 12:01:44			☐	
C:\images\IMG_0309.JPG	Apple	iPhone 5s	2014.05.11 19:17:44	52.3721972222222	4.89758611111111	☐	
C:\images\IMG_0317.JPG	Apple	iPhone 4S	2013.12.08 17:08:04	45.335425	14.3090333333333	☐	
C:\images\20090704-In...						☐	
C:\images\IMG_0444.JPG	Apple	iPhone 5s	2015.09.27 16:42:16	45.8174361111111	15.9762166666667	☐	
C:\images\IMG_1747.JPG	Apple	iPhone 5s	2015.01.02 13:04:15	39.7225416666667	22.6012916666667	☐	
C:\images\img_1771.jpg	Canon	Canon PowerShot S40	2003.12.14 12:01:44			☐	
C:\images\IMG_0482.JPG	Apple	iPhone 5s	2014.06.06 18:19:38	48.3541833333333	11.7908944444444	☐	
C:\images\IMG_0615.JPG						☐	
C:\images\IMG_2100.JPG	Apple	iPhone 5s	2015.01.03 15:15:27	40.555625	22.9929305555556	☐	

Figure 7: EXIF Analysis interface showing information for multiple artifacts in the same table

Upon selection of a data row, the EXIF details of the corresponding artifact are displayed on the right panel along with its associated geolocation, if any, on a map while by clicking on the pinned location a smaller version of the image is displayed (see Figure 9). The simple, but useable, interface allows the association of retained data, assisting in the understanding of the scene by correlating attributes of different artifacts.

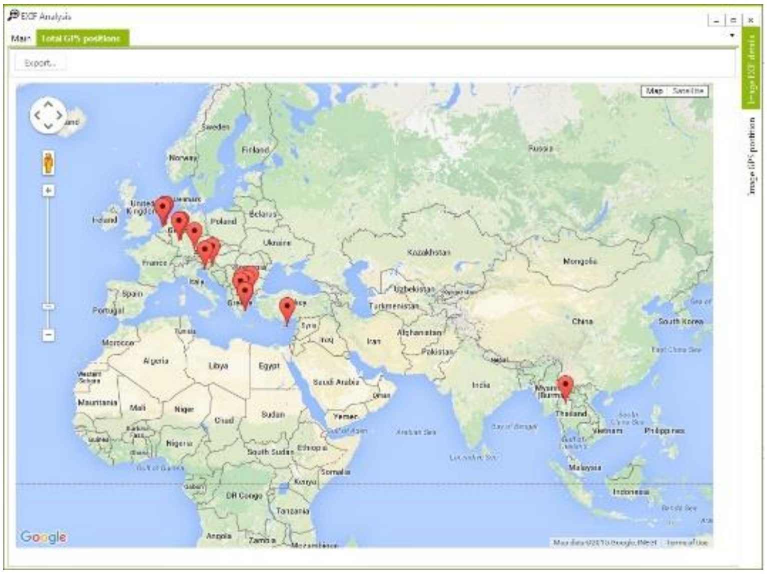
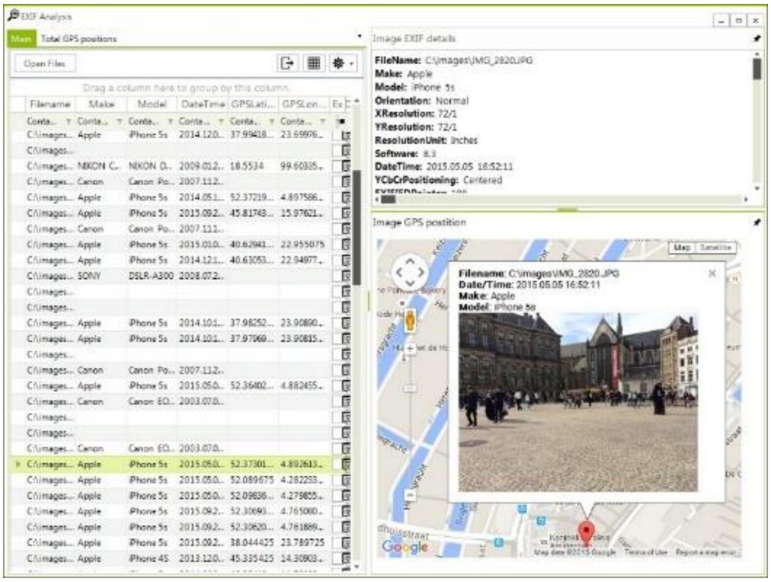


Figure 8: Mapping of artifacts' geolocation

Multiple preconfigured filters as well as custom filters can, also, be applied on any metadata field allowing visualization based on categories of interest. As an additional feature, the tool supports grouping of images according any field (i.e. group artifacts according to their make, model, location etc.). An example is shown in Figure 5.



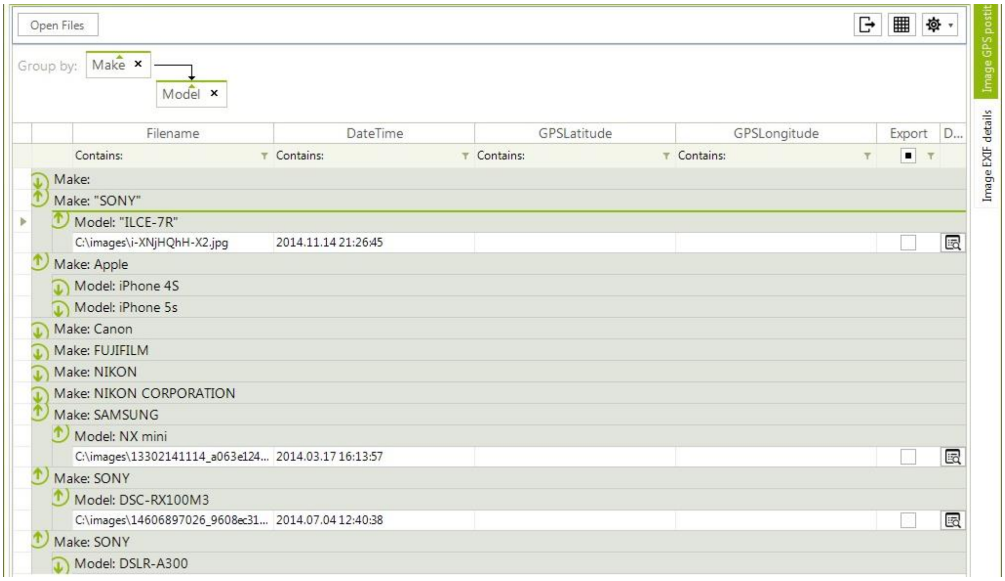
Filename	Make	Model	DateTime	GPSLati...	GPSLon...
Conta...	Conta...	Conta...	Conta...	Conta...	Conta...
C:\images...	Apple	iPhone 5s	2014.120...	37.99418...	23.69976...
C:\images...					
C:\images...	NIKON C...	NIKON D...	2009.012...	18.5534	99.60035...
C:\images...	Canon	Canon Po...	2007.112...		
C:\images...	Apple	iPhone 5s	2014.051...	52.37219...	4.897586...
C:\images...	Apple	iPhone 5s	2015.092...	45.81743...	15.97621...
C:\images...	Canon	Canon Po...	2007.111...		
C:\images...	Apple	iPhone 5s	2015.010...	40.62941...	22.955075
C:\images...	Apple	iPhone 5s	2014.121...	40.63053...	22.94977...
C:\images...	SONY	DSLR-A300	2008.072...		
C:\images...					
C:\images...					
C:\images...	Apple	iPhone 5s	2014.101...	37.98252...	23.90890...
C:\images...	Apple	iPhone 5s	2014.101...	37.97969...	23.90815...
C:\images...					
C:\images...	Canon	Canon Po...	2007.112...		
C:\images...	Apple	iPhone 5s	2015.050...	52.36402...	4.882455...
C:\images...	Canon	Canon EO...	2003.070...		
C:\images...					
C:\images...					
C:\images...	Canon	Canon EO...	2003.070...		
C:\images...	Apple	iPhone 5s	2015.050...	52.37301...	4.892613...
C:\images...	Apple	iPhone 5s	2015.050...	52.089675	4.282233...
C:\images...	Apple	iPhone 5s	2015.050...	52.09836...	4.279855...
C:\images...	Apple	iPhone 5s	2015.092...	52.30693...	4.765080...
C:\images...	Apple	iPhone 5s	2015.092...	52.30620...	4.761889...
C:\images...	Apple	iPhone 5s	2015.092...	38.044425	23.789725
C:\images...	Apple	iPhone 4S	2013.120...	45.335425	14.30903...

FileName: C:\images\IMG_2820.JPG
Make: Apple
Model: iPhone 5s
Orientation: Normal
XResolution: 72/1
YResolution: 72/1
ResolutionUnit: Inches
Software: 8.3
DateTime: 2015.05.05 16:52:11
YCbCrPositioning: Centered

Filename: C:\images\IMG_2820.JPG
Date/Time: 2015.05.05 16:52:11
Make: Apple
Model: iPhone 5s

Figure 9: View of the interface upon the selection of an artifact

The interface also allows exporting entries of interest, into an HTML file. The resulting report includes a geographical map clearly indicating the location of each artifact as well as a table with all metadata for future analysis and evidence preservation, while the actual images are also replicated in a separate folder. During this operation the user is capable of determining various report parameters like the size of the map and fields to be extracted.



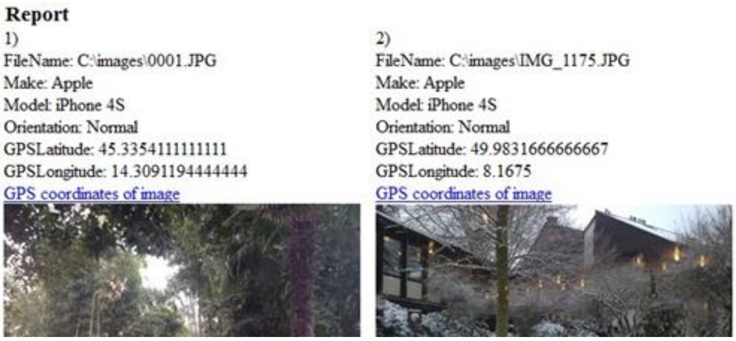
Open Files

Group by: Make × Model ×

Filename	DateTime	GPSLatitude	GPSLongitude	Export	D...
Contains:	Contains:	Contains:	Contains:	■	
Make:					
Make: "SONY"					
Model: "ILCE-7R"					
C:\images\i-XNjHQhH-X2.jpg	2014.11.14 21:26:45			□	
Make: Apple					
Model: iPhone 4S					
Model: iPhone 5s					
Make: Canon					
Make: FUJIFILM					
Make: NIKON					
Make: NIKON CORPORATION					
Make: SAMSUNG					
Model: NX mini					
C:\images\13302141114_a063e124...	2014.03.17 16:13:57			□	
Make: SONY					
Model: DSC-RX100M3					
C:\images\14606897026_9608ec31...	2014.07.04 12:40:38			□	
Make: SONY					
Model: DSLR-A300					

Figure 10: Application of filters on meta-information

Finally, a forensic report, another type of report, containing selected images can be created and customized upon the specific requirements of the DFE. More analytically, the DFE has the ability to select which EXIF details to be exported and in which order; part of a report is shown in Figure 11. This is done to accommodate the need of exporting only desirable fields while the tool supports exports in HTML and CSV format. Upon the generation of a report, the EXIF analysis tool automatically saves user preferences future use minimizing configuration times.



Report

1)
FileName: C:\images\0001.JPG
Make: Apple
Model: iPhone 4S
Orientation: Normal
GPSLatitude: 45.3354111111111
GPSLongitude: 14.3091194444444
GPS coordinates of image

2)
FileName: C:\images\IMG_1175.JPG
Make: Apple
Model: iPhone 4S
Orientation: Normal
GPSLatitude: 49.9831666666667
GPSLongitude: 8.1675
GPS coordinates of image

Figure 11: Example of HTML Report with associated coordinates

5 Discussion & Future Advances

A realistic scenario for a DFE examiner is, provided a number of artifacts, question the authenticity of the material, verify the capturing source and then to reconstruct the depicted scene so as to be able to draw conclusions. Due to the fact that the serial number of the capturing machine is not always recorded as meta-information, verification of images' source is done according to brand and model information, in conjunction with information captured in other fields (i.e. geo-location, timestamp etc.). The analysis process involves varying degrees of uncertainty.

A common forensic case for authorities in Cyprus is the investigation of child pornographic cases for which the utilization of images is rather critical. The use of the EXIF analysis tool in such cases will assist in determining if images have been captured by cameras of the same brand, model or even type, as well as the geolocation of the event. In some cases, the examiner may even manage to correlate the identity of the person behind an identified camera with images captured in different geolocations, providing evidence that a specific individual has exploited a number of children at different moments in time. The aforementioned case serves as an example application of the proposed tool but it does not limit by any sense its usability in other cases like the investigation of content admitting terrorism events, unauthorized access to restricted areas/regions (i.e. military bases, archaeological sites), capturing of critical infrastructures with malicious intentions and general crime investigation.

As a future development the use of MD5 signatures could increase even more the acceptance of the tool by DFEs as the use of signatures enhances the credibility and trustworthness of the tool additionally, the devlopment of a mechanism for the reconstruction of scenarios from identified devices considering the order in which events have occurred as well as the interval between records. .

6 Conclusions

Digital forensic examination known as a time consuming and challenging task where the DFE seeks for answers to investigation questions through analysis of digital artifacts. Inference of conclusions requires combination of contextual temporal and spatial information for the reconstruction of the scene. Despite advances in the timely processing of vast amounts of data current digital forensics tools lack intuitive interfaces. A new DFT developed by ADITESS for the automatic extraction of embedded information present in the form of EXIF metadata in images has been presented. EXIF analysis has been designed based on end-user requirements and assists the examination process with the provision of temporal, spatial and contextual information through a user friendly and useable interface.

Acknowledgements

The work presented in this paper received funding from the European Union ISEC programme (HOME/2013/ISEC) entitled Cybercrime Center of Excellence Network under grant agreement number HOME/2013/ISEC/AG/INT/4000005229.

References

[1] G. Palmer, "A road map for digital forensics research-report from the first Digital Forensics Research Workshop (DFRWS)," *Utica, New York*, 2001.

[2] G. L. Palmer, "Forensic analysis in the digital world," *Int. J. Digit. Evid.*, vol. 1, no. 1, pp. 1–6, 2002.

[3] D. R. Kamble and N. Jain, "DIGITAL FORENSIC TOOLS: A COMPARATIVE APPROACH."

[4] J. F. Nunamaker Jr, N. C. Romano Jr, and R. O. Briggs, "A framework for collaboration and knowledge management," in *System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on*, 2001, p. 12–pp.

[5] S. L. Garfinkel, "Digital forensics research: The next 10 years," *Digit. Investig.*, vol. 7, pp. S64–S73, 2010.

[6] Technical Standardization Committee on AV & IT Storage Systems and Equipment, "Exchangeable image file format for digital still cameras: EXIF version 2.2." Standard of Japan Electronics and Information Technology Industries Association, 2002.

[7] "PhotoDNA Cloud Service." [Online]. Available: https://www.microsoft.com/en-us/PhotoDNA/Default. [Accessed: 05-Sep-2015].

Advanced Analytics with SAP HANA

Jordan Jordanov, Bogdan Vatkov, Dimitar Angelov

SAP Labs Bulgaria, EOOD

Blvd. Tzar Boris III 136A, Sofia, Bulgaria

jordan.jordanov@sap.com; bogdan.vatkov@sap.com; dimitar.angelov@sap.com;

Abstract: SAP HANA is an in-memory computing platform that has completely transformed the relational database industry. It combines database, application processing, and integration services on a single platform. It provides libraries for predictive, planning, text processing, spatial, and business analytics. In this paper we will have a look at a selected subset of the data processing engines that deliver advances analytics capabilities which go beyond the classical OLAP analytics.

Keywords: In-memory computing, predictive analytics, text analysis, text mining, sentiment analysis, spatial analysis.

1 Introduction

SAP HANA was created back in 2008 as a result of the collaboration between SAP SE, the Hasso Plattner Institute and the Stanford University. In 2010 it was introduced as a product to the market, in 2012 became available in the Cloud and in 2013 it already powered the SAP Business Suite – the world's leading integrated portfolio of software applications for large and mid-sized businesses. Technology wise SAP HANA started as an In-Memory RDBMS with real-time OLAP capabilities but nowadays it is a rich data platform with large set of data processing engines covering specific domains like predictive analytics, text analysis, graph processing, spatial processing, complex event processing, and others – all based on in-memory computing principles.

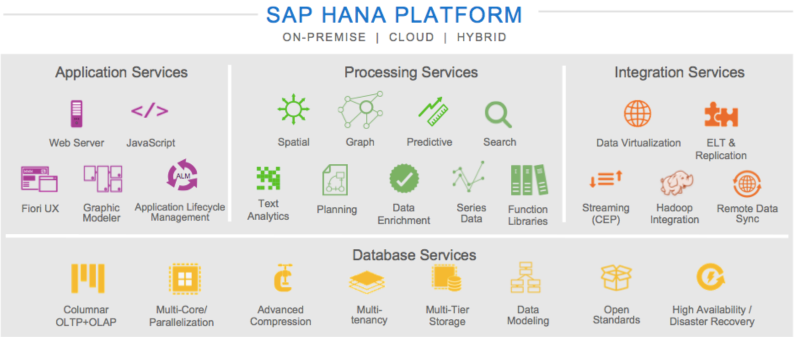


Figure 1: SAP HANA Platform

2 Parallel Data Flow Computing

Apart from the OLAP Engine which provides classical BI functions like drill-in, roll-up, slice and dice, SAP HANA also features a generic parallel computing engine, also known as the Calculation Engine, which allows arbitrary sophisticated data processing flows to be modeled and executed in an optimal way.

To natively take advantage of massively parallel multicore processors, SAP HANA manages the SQL processing instructions into an optimized model that allows parallel execution and scales incredibly well with the number of cores. The optimization includes partitioning the data in sections for which the calculations can be executed in parallel. Execution can also be scaled across multiple hosts.

The data processing execution control is organized by the so called calculation models. A calculation model is a directed acyclic graph (DAG) with arrows representing data flows and nodes that represent operations. This approach and the exclusion of loops and recursion enable automatic massive parallelism of the processing. The easiest way to think of calculation models is to see them as data flow graphs, where the modeler defines data sources as inputs and different operations (join, aggregation, projection, and so on) on top of them for data manipulation. The calculation engine automatically breaks up a model into operations that can be processed in parallel (model optimizer). These operations are passed to the database optimizer, which determines the best plan for accessing row or column stores, leveraging cost-based optimizations and database statistics.

Developers can leverage the Calculation Engine by composing SQLScript programs. SQLScript is a collection of extensions to the Structured Query Language (SQL). The extensions are: Data extension, which allows the definition of table types without corresponding tables; Functional extension, which allows definitions of (side-effect free) functions which can be used to express and encapsulate complex data flows; Procedural extension, which provides imperative constructs executed in the context of the database process. Conceptually SQLScript is related to stored procedures as defined in the SQL standard, but SQLScript is also designed to provide superior optimization possibilities.

3 Predictive Analytics

As we already outlined SQLScript includes enhanced control-flow capabilities and lets developers define complex application logic inside database procedures. However, it is difficult to describe predictive analysis logic with procedures.

For example, an application may need to perform a cluster analysis in a huge customer table with billions of records. It is impossible to implement the analysis in a procedure using the simple classic K-means algorithms, or with more complicated algorithms in the data-mining area. Transferring large tables to the application server to perform the K-means calculation is also costly. That is why predictive analysis is delivered through functions. The Predictive Analysis Library (PAL) defines functions that can be called from within SQLScript procedures to perform analytic algorithms. PAL includes large set of classic and universal predictive analysis algorithms distributed in several data-mining categories:

- Association Analysis
  - Apriori
  - Apriori Lite
  - FP-Growth
  - KORD –Top K Rule Discovery
- Classification Analysis
  - CART
  - C4.5 Decision Tree Analysis
  - CHAID Decision Tree Analysis
  - K Nearest Neighbor
  - Logistic Regression
  - Back-Propagation (Neural Network)
  - Naïve Bayes
  - Support Vector Machine
  - Confusion Matrix
  - Parameter Selection & Model Evaluation
- Regression
  - Multiple Linear Regression
  - Polynomial Regression
  - Exponential Regression
  - Bi-Variate Geometric Regression
  - Bi-Variate Logarithmic Regression
- Cluster Analysis
  - ABC Classification
  - DBSCAN
  - K-Means
  - K-Medoid Clustering
  - K-Medians
  - Kohonen Self Organized Maps
  - Agglomerate Hierarchical
  - Affinity Propagation
  - Gaussian Mixture Model
  - Latent Dirichlet Allocation (LDA)
- Time Series Analysis
  - Single Exponential Smoothing
  - Double Exponential Smoothing
  - Triple Exponential Smoothing
  - Forecast Smoothing
  - ARIMA / Seasonal ARIMA
  - Brown Exponential Smoothing
  - Croston Method
  - Forecast Accuracy Measure
  - Linear Regression with Damped Trend and Seasonal Adjust
  - Test for White Noise, Trend, Seasonality
- Probability Distribution
  - Distribution Fit
  - Cumulative Distribution Function
  - Quantile Function
- Outlier Detection
  - Inter-Quartile Range Test (Tukey’s Test)
  - Variance Test
  - Anomaly Detection
  - Grubbs Outlier Test
- Link Prediction
  - Common Neighbors
  - Jaccard’sCoefficient
  - Adamic/Adar
  - Katzβ
- Data Preparation
  - Sampling
  - Random Distribution Sampling
  - Binning
  - Scaling
  - Partitioning
  - Principal Component Analysis (PCA)
- Statistic Functions (Univariate)
  - Mean, Median, Variance, Standard Deviation
  - Kurtosis
  - Skewness
- Statistic Functions (Multivariate)
  - Covariance Matrix
  - Pearson Correlations Matrix
  - Chi-squared Tests
    - Test of Quality of Fit
    - Test of Independence
  - F-test (variance equal test)
- Miscellaneous
  - Weighted Scores Table
  - Substitute Missing Values

4 Text Analysis

Text analysis in SAP HANA is a set of natural-language processing capabilities based on linguistic, statistical and machine-learning algorithms that infers models and structures out of textual sources in multiple languages. This technology forms the foundation for advanced text processing for a range of applications including text search, business intelligence or exploratory data analysis. There are three major areas of text analysis functionality: Linguistic Analysis, Entity Extraction, Fact Extraction and Grammatical Role Analysis.

Linguistic Analysis is the most fundamental form of text analysis - tokenization, stemming and part-of-speech tagging. These operations allow for optimized full-text index building because they guarantee both high precision and recall.

Entity Extraction is the identification of named entities (persons, organizations etc.), which eliminates the 'noise' in textual data by highlighting salient information. This process transforms unstructured text into structured information. Extracting entities from unstructured text tells us what the text is about – the people, organizations, places, and other parties described in the document. The extraction process involves processing and analyzing text, finding entities of interest, assigning them to the appropriate type, and presenting this metadata in a standard format, including the entity's character offset and length in the document, and other attributes. The extraction process can extract entities using lists of specific named entities. Entities are often proper names, such as the names of specific and unique people, organizations, or places. Other specified entity types include currency amounts and dates, among others. Each entity is defined as a pairing of a name and its type. For example:

Canada/COUNTRY

John Paul/PERSON

General Motors Corporation/ORGANIZATION/COMMERCIAL

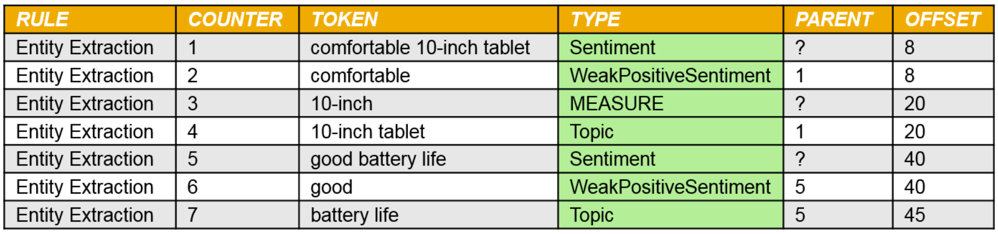
Entity types play a crucial role in the definition of an entity. Entity types are used to classify entities extracted from documents and entities stored in a dictionary. There is an extensive set of predefined entity types and you can optionally enhance the extraction process by using dictionaries.

Fact Extraction is a higher-level semantic processing that links entities as "facts" in domain-specific applications; it includes detection of sentiments expressed about something, public-sector-specific information and enterprise facts. There are several configurations for fact extraction supported:

- Core – extracts entities of interest from unstructured text, such as people, organizations, or places mentioned. In most use cases, this option is sufficient.

- Enterprise – includes a set of entity types and rules for extracting information about organizations, such as management changes, product releases, mergers, acquisitions, and affiliations.
- Public Sector – includes a set of entity types and rules for extracting information about events, persons, organizations, and their relationships, specifically oriented towards security-related events.
- Voice of the Customer – includes a set of entity types and rules that address requirements for extracting customer sentiments and requests. You can use this content to retrieve specific information about your customers' needs and perceptions when processing and analyzing text. This configuration involves complex linguistic analysis and pattern matching that includes processing parts of speech, syntactic patterns, negation, and so on, to identify the patterns to be extracted. The keyword dictionaries used to identify and classify sentiments can also be customized, if needed, for specific applications.

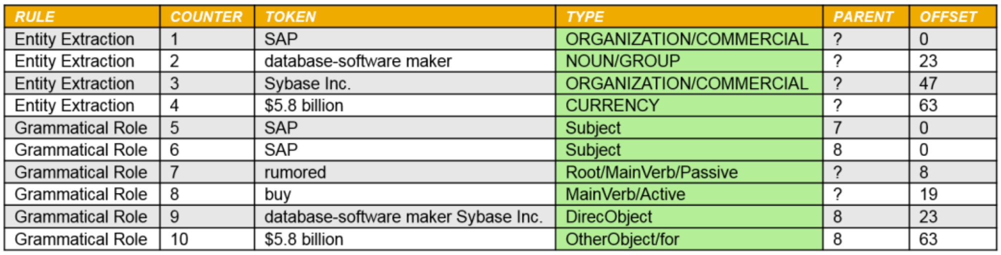
If we run a Voice of the Customer analysis of “It is a comfortable 10-inch tablet with good battery life.” we will get sentiments as well as topic reported like the ones on figure 2.



RULE	COUNTER	TOKEN	TYPE	PARENT	OFFSET
Entity Extraction	1	comfortable 10-inch tablet	Sentiment	?	8
Entity Extraction	2	comfortable	WeakPositiveSentiment	1	8
Entity Extraction	3	10-inch	MEASURE	?	20
Entity Extraction	4	10-inch tablet	Topic	1	20
Entity Extraction	5	good battery life	Sentiment	?	40
Entity Extraction	6	good	WeakPositiveSentiment	5	40
Entity Extraction	7	battery life	Topic	5	45

Figure 2: Entity Extraction with Sentiment Analysis for a sample input

The Grammatical Role Analysis capability identifies syntactic relationships between elements of a sentence in the form of subject–verb–object expressions, commonly known as ‘triples’. If we run Grammatical Role Analysis over the example text: “SAP was rumored to buy database-software maker Sybase Inc. for $5.8 billion.” we will get extracted subjects, passive and active verbs, objects like the ones on the figure 3.



RULE	COUNTER	TOKEN	TYPE	PARENT	OFFSET
Entity Extraction	1	SAP	ORGANIZATION/COMMERCIAL	?	0
Entity Extraction	2	database-software maker	NOUN/GROUP	?	23
Entity Extraction	3	Sybase Inc.	ORGANIZATION/COMMERCIAL	?	47
Entity Extraction	4	$5.8 billion	CURRENCY	?	63
Grammatical Role	5	SAP	Subject	7	0
Grammatical Role	6	SAP	Subject	8	0
Grammatical Role	7	rumored	Root/MainVerb/Passive	?	8
Grammatical Role	8	buy	MainVerb/Active	?	19
Grammatical Role	9	database-software maker Sybase Inc.	DirecObject	8	23
Grammatical Role	10	$5.8 billion	OtherObject/for	8	63

Figure 3: Entity Extraction with Grammatical Role Analysis for a sample input

The following data types in the database are enabled for text analysis: TEXT, BINTEXT, NVARCHAR, VARCHAR, NCLOB, CLOB, and BLOB. There are more than 30 languages supported by the Text Analysis engine.

5 Text Mining

Text mining provides functionality that can compare documents by examining the terms used within them. It supports functions to determine the top-ranked related and relevant documents and terms relative to a given reference document set and perform statistical analysis. It also supports methods for categorizing (classifying) documents, given a set of reference documents with predefined categorizations.

Text mining works at the document level – it allows to make semantic determinations about the overall content of documents relative to other documents. This is in contrast to text analysis, which does linguistic analysis and extracts information embedded within each document. They are complementary. Text mining benefits by using the information determined by text analysis. Text Mining provides several types of functions: Initialization, Term-Document functions, Categorization (Classification) and Statistical Analysis.

When text mining is initialized on the reference document content in a given table and column, it creates and retains the term-document matrix and other configuration context needed for text mining functions. Text mining initialization of reference data is normally done in advance when the full-text index is created for the given table and column. An explicit initialization call is also provided for advanced use. Input documents are typically separate from the reference data and are automatically processed by text mining at run time as needed, so initialization does not need to be invoked for them.

Term-Document functions are determining top-ranked related and relevant documents and terms.

Categorization is based on comparing an input document with a reference set of already-classified documents. The reference set documents must contain a suitably complete and representative vocabulary and be sufficiently numerous for each categorization choice. The simplest case is to have a single categorization column that can have a single value. Text mining can also support multiple independent categorization columns of this type. For example, each document could have a subject and a geographical region. Text mining supports categorization using the KNN Classifier. No training (up-front preparation) is required – it uses the reference set documents directly. The KNN Classifier determines the K Nearest Neighbors (most-similar documents) from the reference set and then sums and normalizes their similarities per category value to determine the winning category value. The similarity of two documents is related to how much their vectors point in the same direction. Text mining provides several standard similarity measures – Cosine, Jaccard, Dice (Sørensen) and Overlap.

Statistical Analysis for related and relevant documents and terms can be done using several functions – Correlation Matrix, Principal Component Analysis (dimensionality reduction) and Clustering. Correlation Matrix returns a term or document correlation matrix. The matrix contains the correlation values between each of the returned results. Principal Component Analysis returns factor and rotation values for the specified number of principal components, for each of the returned results. Clustering does hierarchical clustering with the method specified: Complete Linkage, Single Linkage, Average Distance Within, Average Distance Between, or Ward's Method. It returns the clustering level, left value, and right value for each of the returned results.

The text mining functions are offered via the SAP HANA Text Mining XS JavaScript API and via SQL.

6 Spatial Processing

Column-oriented data structures and in-memory computing have developed into powerful components of today's enterprise applications. While the focus of these developments has primarily been on analyzing sales data, the potential for using these technologies to analyze geographic information is significant. Support for the processing of spatial data represents a key evolution in SAP HANA.

To deliver vastly improved performance and results in everything from modeling and storage to analysis and presentation of spatial data, SAP HANA includes a multilayered spatial engine and supports spatial columns, spatial access methods, and spatial reference systems.

Spatial data is data that describes the position, shape, and orientation of objects in a defined space. Spatial data is represented as 2D geometries in the form of points, line strings, and polygons. Two common operations performed on spatial data are calculating the distance between geometries and determining the union or intersection of multiple objects. These calculations are performed using predicates such as intersects, contains and crosses.

Spatial data support lets application developers associate spatial information with their data. For example, a table representing companies could store the location of the company as a point, or store the delivery area for the company as a polygon. Here is an example on how to search for a Company based on its delivery area definition and a particular point of interest:

```
SELECT * FROM CompanyLocations WHERE DeliveryArea.ST_Contains(NEW
ST_Point('POINT ( 23.321868 42.697708 )')) = 1
```

ST_POINT is composed of longitude and latitude.

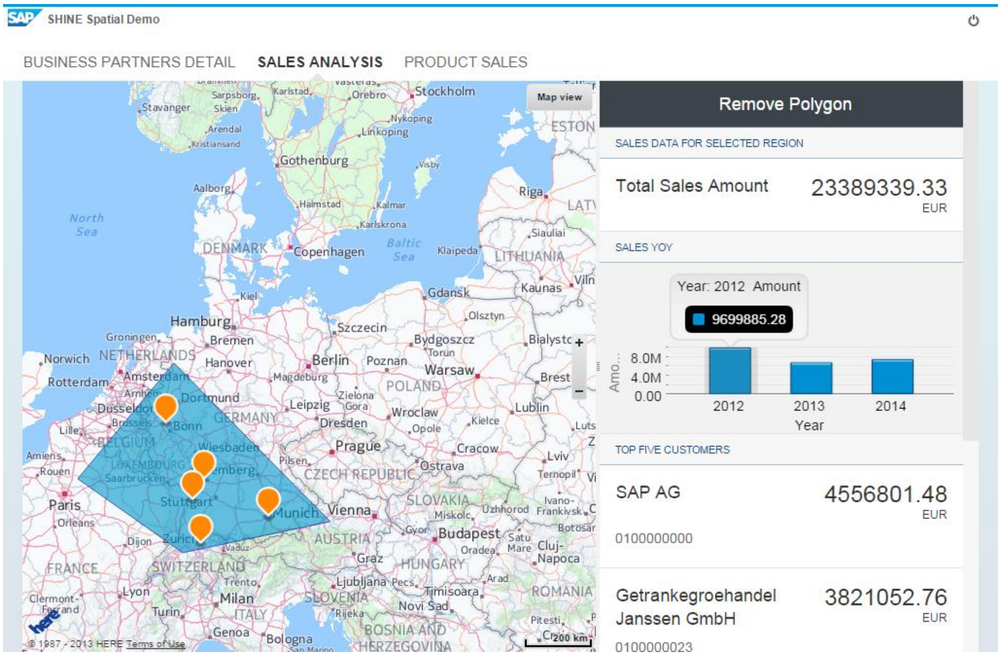
The following spatial data types are supported: Geometries, Points, Multipoints, Linestrings, Multilinestrings, Polygons, Multipolygons, Circularstrings

Spatial predicates are implemented as member functions that return 0 or 1. To test a spatial predicate, your query should compare the result of the function to 1 or 0 using the = or <> operator. For example:

*SELECT * FROM SpatialShapes WHERE shape.ST_IsEmpty() = 0;*

Supported import and export formats are: Well Known Text (WKT), Extended Well Known Text (EWKT), Well Known Binary (WKB), Extended Well Known Binary (EWKB), ESRI shapefiles, GeoJSON and Scalable Vector Graphic (SVG) files.

Geocoding is also supported. It is the process of converting address information into geographic locations. Since current business data is associated with address information without explicit geo-location information, geocoding is crucial to enable geographic analysis of this data. The geocoding itself is provided by an external provider such as Nokia NAVTEQ or TOMTOM. The geocoding processing is very similar to text processing, with the difference that multiple columns are used as an input, only asynchronous processing is supported, and the geocode column (the column which holds the converted addresses) is visible. The GEOCODE column is visible for searches, but must not be altered or dropped.



SALES DATA FOR SELECTED REGION

Total Sales Amount 23389339.33 EUR

SALES YOY

TOP FIVE CUSTOMERS

SAP AG 4556801.48 EUR
0100000000

Getrankegroehandel Janssen GmbH 3821052.76 EUR
0100000023

Figure 4: Total sales for a custom geographical region using spatial analysis

Let’s have a look at a simple example “Sales Analysis” scenario. We can use spatial processing to perform simple analysis over a particular geographical region. Using boundaries selected on a map we derive a polygon definition. Based on this polygon definition, we can calculate the business partners that lie within this area with a select statement, for example:

SELECT STREET, NEW ST_Point(LATITUDE, LONGITUDE).ST_Within(NEW ST_Polygon('Polygon((0.0 0.0,90.0 0.0, 90.0 90.0, 0.0 90.0, 0.0 0.0))')) AS IS_INSIDE from MD.Addresses

SAP HANA SQLScript has been extended to support the Spatial SQL MM Standards for storing and accessing geospatial data. SAP HANA also delivers Nokia mapping services as part of the HANA Spatial features.

7 Conclusions

With the ever increasing variety of data sources modern Big Data processing solutions are expected to offer advanced analytics capabilities like Text Analysis, Text Mining and Spatial Processing to extract valuable information. By combining the classical in-memory RDBMS and OLAP capabilities with advanced analytics features like Predictive Analysis, Text Analysis, Text Mining and Spatial Processing SAP HANA enables complex processing of large and diverse datasets to derive deeper insights like never before.

References

[1] Franz Faerber, Sang Kyun Cha, Juergen Primsch, Christof Bornhoevd, Stefan Sigg, Wolfgang Lehner (2011) SAP HANA Database - Data Management for Modern Business Applications, SIGMOD Record, December 2011 (Vol. 40, No. 4), pages X-Y.

[2] SAP HANA Predictive Analysis Library (PAL), SAP HANA Platform SPS 10, Document Version: 1.1 – 2015-08-21

[3] http://go.sap.com/solution/in-memory-platform.html

[4] https://blogs.saphana.com/2015/06/16/new-sap-hana-sps10/

[5] https://blogs.saphana.com/2013/09/19/spatial-processing-with-sap-hana/

[6] https://github.com/SAP/hana-shine/tree/master/Tutorials/SHINE-SpatialScenario

http://help.sap.com/hana_options_adp

Big Data in Astroinformatics - Compression of Scanned Astronomical Photographic Plates

Vasil Kolev

IICT - BAS,

Sofia, Bulgaria

e-mail: kolev_acad@abv.bg

Abstract: Construction of SPPs databases and SVD image compression algorithm are considered. Some examples of compression with different plates are shown.

Keywords: SVD, Database, Scanned astronomical photographical plates,

1 Introduction

The newly born area of Astroinformatics has emerged as an interdisciplinary area from Astronomy and modern information and communication technologies, based on the modern Internet developments. As a truly interdisciplinary area Astroinformatics has arisen from the need of information and communication technology (ICT) methods for preservation and exploitation of the scientific, cultural and historic heritage of astronomical observations. Practically, before the application of the charge-coupled devices (CCD) and orbital multi-wave telescopes the only source of information in Astronomy were the astronomical photographic observations into astronomical plates. The total information recorded on astronomical plates is estimated about 1PB (=106GB=109MB=1015Bytes) provided the information is in computer readable format. It suggests that after the plate inventory and proposed plate catalogue the plates themselves are digitized and online access to the plate image is provided. With the new development of the information and communication technologies and Internet development including the digitization and loss- lossless compression technologies of this problem can be solved completely in the near future.

One of the main problems is that the scanned photographic astronomical images have a huge volume (for example: the volume of one scanned astronomical plate is from 100 MB up to 2,000 MB). The access to such volume of information, as well as its analysis, is now difficult and slow task. On the other hand it is not needed for majority of tasks for analyzing the entire information available on the plate. That is why it is very essentially the digital information to be compressed in such way that the plate information to be stored and retrieved in suitable way after reduction and standardization for usage from many different users. For these purposes exceptionally useful are the methods for compression, analysis and image proceedings, developed intensively during the last 10-15 years. Their development is due mainly to the Internet progress, modern GRID technologies and GRID-based virtual portals, in particular virtual observatories.

The astronomical photographic plates (APP) obtained with a particular telescope at a particular observational site and stored at a definite place constitute the so-called plate database archive where the number ranges between several dozen to more than 100 000. Only a small number of archives have more than 10 000 plates. The largest European photographic plate database is in Sonneberg Observatory with 250 000 plates. The WIDE-FIELD PLATE DATABASE (WFPDB) is present at the Sofia Sky Archive Data Center is shown in [1] with 464,373 SPPs. The scanning process of APP includes plate images with optimal low- and high-resolution scans tabulated of Table 1 [2] and form of scanned astronomical photographic plates (SAPP) database. It is a virtual unique instrument-telescope, which allows from SAPPs searching for necessary plate index information in the available world plate vaults for the period of the last 130 years - the era of photographic astronomy.

Astronomical observations produce a large amount and a large variety of data. Many measurements of some old plates were experimented and SAPPs - however show important defects. The lifetime of archived data is nowadays a problem, since magnetic tapes have only a short lifetime (a few years), and the large amount of data involved discourages regular duplications. The SAPPs are very large data sets lead to moving to a new computer should reading data stored on the presently used media. Example, the size of some scanned images of SAPP saved in FITS files are shown in Table 2 where astronomical data usually are processing in next steps:

- First, acquired from a receptor installed on an astronomical instrument, via a command and control process; the output of the detector is called raw data (e.g. the astronomical photographic plate).

- The next process consists in detecting anomalies and removing known instrumental biases by means of calibration adjustments, and leads to calibrated or edited data.

- Last and important process is how to reduce and analyzed data expressed in physical units (e.g. fluxes as a function of wavelength). This step makes use of more sophisticated image or data processing (filtering, contrast enhancement, model fitting, etc), and frequently implies a significant reduction of the data size as compression of the scanned images.

TABLE I

Summarizes the number of SAPPs in Europe with flatbed scanners with high (20 microns/pixel), and low resolution;

Observatory	Resolution	
	High	Low
Sonneberg	215 000	
Pulkovo	40 000	
Hamburg	8 000	
Vatican	5433	
Tautenburg	4228	
Asiago	4000	
Buyrakan	2000	
Potsdam	1500	
Bamberg	1000	2000
Bucharest		5000
Belgrade		2000
Sofia		1500
Moscow-Zvenigorod		1000

TABLE II

Size of *.FITS files of the SAPP images with high resolution scans

Telescope Identifier	EPSON Flatbed Scanner	Plate Size (cm.)	High Resolution (dpi)	FITS file (MB)
ESO100	Expression 10000 XL	30×30	2400	1670
ROZ200	Expression 10000 XL	30×30	1600	664
POT032	Expression 10000 XL	16×16	2400	440
BAM10C	Perfection V750 Photo	16×16	2400	430
POT032	Perfection V700 Photo	16×16	2400	412
BON034	Perfection V750 Photo	8×9	2400	126

2 Astronomical Big Data

The astronomical photographic plates constitute glass coated on one side with a dry emulsion of silver bromide. As light detectors and media for information storage, they replaced visual astronomical observations and marked the epoch of photographic astronomy that began in the 1870s and lasted for more than 130 years. Many astronomical discoveries were made on the basis of photographic plates used in the observations. Moreover, APPs stored in observatories or relevant institutions continue to be used for different astronomical tasks, especially such ones which need long time series observations as the APPs are typical examples of scientific heritage in need of preservation for future use.

Astronomical Big Data is usually defined in terms of the three V's – *Volume, Velocity* and *Variety* [3]:

1. *Volume, Velocity* – they are obvious (large data sets and data streaming or data for stars "in motion", example comets),

2. *Variety* – this is relates to the diversity of data type – this can be the considered to relate to the myriad of different objects, wavelength range/messenger type recorded, and data format as usually is used FITS format.

3. *Veracity* – this is important for analytics, and for astronomers dealing with measurements that are inherently uncertain or possibly corrupt.

The new generation of radio telescopes e.g. LOFAR [4] are often considered as good examples of Astronomical Big Data routinely produce data sets of ~ 100 TB per day, and new facilities such as the SKA [5, 6] generate object catalogues that contain billions of sources [7]. Combining this information requires a new approach to large-scale data analysis where Astronomical Big Data can also be applicable to the astronomy case, example a analytics and visualization of scanned astronomical photographical plates. Progress in data visualization has been at a virtual standstill but data analytics - new visualization approaches, automatic feature extraction tools are constructed. Many of the existing approaches to image processing are still sequential and written for single CPU cores using single threads. Now day we can use multi-threaded software for multiple CPU cores. But exist limitation - network bandwidth which is becomes a limiting factor when large volumes of data are processing and transmutation. This means that single server multi-core CPU systems may not provide enough of a performance to processing rates. The SDSS [8], which is currently in operation using a 2.5 meter telescope, reported a maximum data capture rate in the region of 200GB per day at present [9].

3 Present Standard Astronomical Image Formats

FITS Formats - The FITS format (the Flexible Image Transport System) is a standard astronomical image format endorsed by the International Astronomical Union (IAU) and NASA, which was approved in 1981. There exist libraries supported by NASA to access and manipulate the FITS file format of many languages with interfaces allowing reduction software - Python, Java, Perl, MATLAB and C++.

IRAF - Image Reduction and Analysis Facility is a standard application used throughout the astronomical community which has been in development since the mid to early

1980s [10], and is a Linux based software package. It provides the closest thing to a standard for data reduction and analysis within much of the astronomy community.

NHPPS - The NHPPS is a python based pipeline which can operate with local processing clusters of software nodes based on the OPUS system [11]. The NHPPS uses the blackboard architecture for communication across a multi-node distributed environment which is a multi-queue based system.

ESO: Common Pipeline Library (CPL)

CPL is Linux based operating systems which implementations of existed within various ESO instruments to astronomical data compression as require faster processors to run pipelines and does not provide multi-threaded support [12].

4 Computing of SVD for SAPP

4.1 General Theory

As an illustration we consider the SVD compression of the SAPP images. On a computer, the image is simply a matrix **A** with size $m \times n$ denoting pixel colors. Typically, such matrices can be compressed by low-rank matrices. Instead of storing the $m \times n$ entries of the matrix **A**, one need only store the $k(\mathrm{m}+\mathrm{n})+k$ numbers in the sum:

$$\mathbf{A}_k = \sum_{j=1}^{k} \sigma_j \mathrm{u}_j \mathrm{v}_\mathrm{j}^\mathrm{T}$$

where σ_j are the singular values, k - rank of the matrix, and the matrices u_j and v_j are unitary. When k<<min(m,n) this can make for a significant improvement, though modern image compression protocols use more sophisticated approaches. The SVD is stable, small perturbation in **A** correspondent to small perturbation in the matrix of singular values σ_j and conversely. The reduced rank approximations based on the SVD are very similar in intent. However, SVD captures the best possible basis vectors for the particular data observed, rather than using one standard basis for all cases.

4.2 SVD for SPP database compresssion

Construction of DATABASE of compressed SAPP is show of Fig.1. First, the APPs have to scanned, and after that can be applied any image processing. Scanning process is difficult and time-consummation stage about some hours because of astronomical photographical plate images used for astronomical tasks are made at an optimal high resolution of 1600 (or 2400) dpi in FITS format, see Table 2. The scans are usually available upon request, thus the copyright of the observatories is protected. For bad quality plates only previews are needed. The systematic plate scanning takes considerable funds and gives a huge volume of scan data, which have to be stored. One possible solution of this problem is to image compression of the digitized plates.

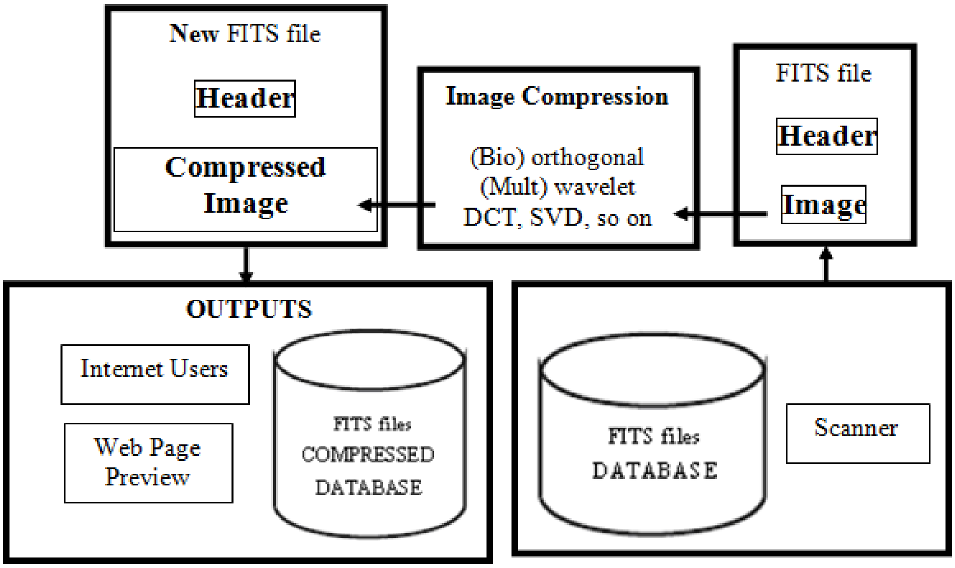


Fig.1 Construction of DATABASE of SAPP

If have already SAPP database or can download from I-net, passed to the image processing step. In this step, we applied different type spline, wavelet, curvelet, multiwavelet, DCT, SVD and so on to image compressing, or extract celestial objects, or examine of the preview image, catalog construction or other one. In [13] is propose a slight modification to the SVD, which gives us much better compression than the standard compression DCT process. For computing step we can using follow hardware resourses:

- PC with 1 procesoor
- Server with 1 procesoor
- Server with more procesoors
- Distributed Computing
- Cloud Computing

The follow step is inserting of the compression image and construct compression FITS file together header. In the output we can using compressed FITS file for Web Preview, construction or add on to existed database, or send to any I-net user.

5 SVD Compression of Scanned Photographical Plates

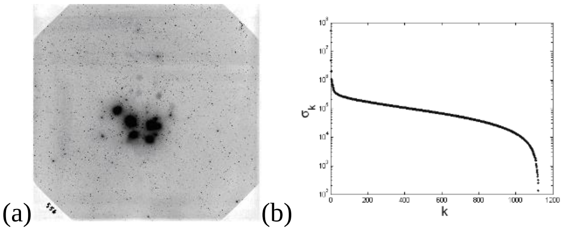


Fig.2 Image of *M45-556p.fits* in the region of the Pleiades stellar cluster

a) Original image of SPP (size 1122x1122), b) Singular values

The singular values obtained from SVD for the scanned plate *M45-556p.fits* in the region of the Pleiades stellar cluster, with image size 1112x1122 pixels and 16bit format is show of Fig.2(a). We see that the singular values decrease near linearly. Though the singular values are very large, $\sigma_1 > 10^7$, there is a relative difference of five orders of magnitude between the smallest and largest singular value show in Fig.2(b). We see that the singular values decrease rapidly.

If all the singular values were roughly the same, we would not expect accurate lossless compression. We can compressed a matrix by adding only the first few terms of the series (Fig. 3).

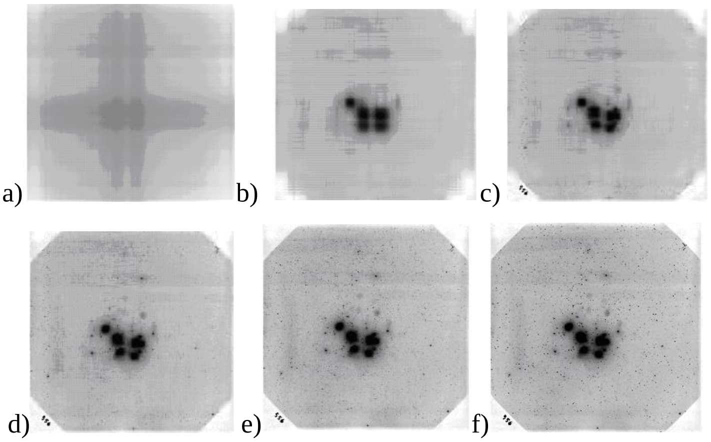
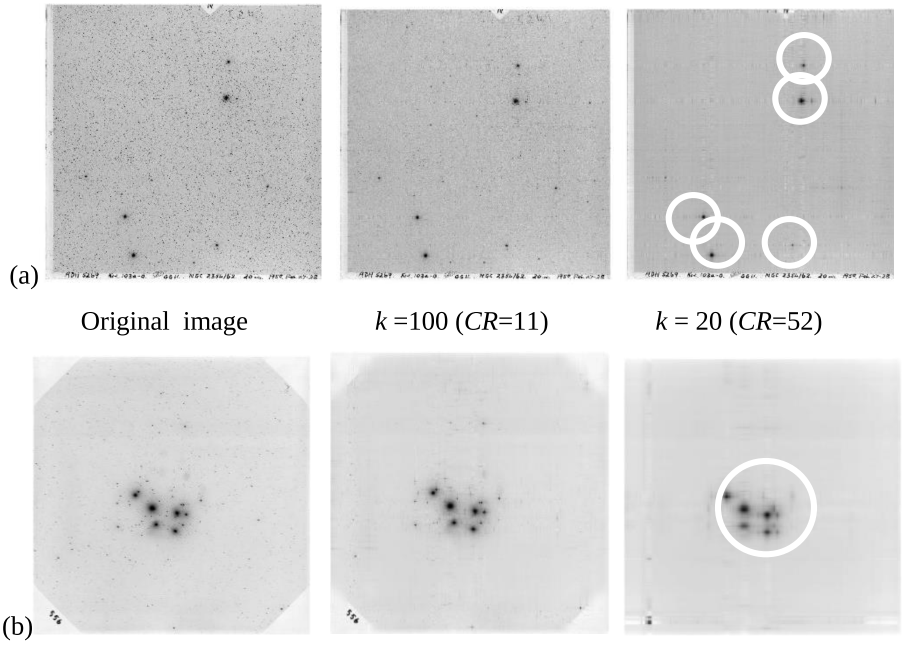


Fig.3 The SVD compression of SPP *M45-556p.fits* in the region of the Pleiades stellar cluster; (a) k=1; (b) k=5 (c) k=12; (d) k=30; (e) k=100; (f) original image

The measure to image compressing is the compression ratio (CR) for SAPP images which is CR = Output image /Input image. Obvious, although CR>50 we recognized main objects as well as show in Fig.4(a). Furthermore, for CR>100 from compressed image Fig.4(b) we easy recognized both main objects and from which SPP is this.



Original image *k* =100 (*CR*=11) *k* = 20 (*CR*=52)

Original image *k*=25 (CR=22) *k*=5 (CR=112)

Fig.4 Compression of the image from SPP; (a) ADH5269.fits with 2048x2048 px; and (b) *M45-556p-1.fits* with 1120x1120 pixels.

Conclusion

In conslusion, for Astronomical Big data great approach to decreaing volume of database is SVD compresion becase of for big CR the inportant image details from SPP are recognnized. This can be use similar as filter and useful for image denoising. The SVD compression is faster from wiener filter processing. Therefore, compressed images do require less computer storage in database and transmission time than the full-rank image.

References

[11] Tsvetkov M., (2006) Wide-Field Plate Database: a Decade of Development, Virtual Observatory: Plate Content Digitization, Archive Mining & Image Sequence Processing, Sofia.

[12] Tsvetkov M., (2012) Tsvetkova K., Kirov N., Technology for Scanning Astronomical Photographic Plates, Serdica J. Computing. vol. 6, pp. 77–88.

[13] Laney D., (2001) Data Management: Controlling Data Volume, Velocity, and Variety *META, Technical report*.

[14] Haarlem M., (2013) Wise M., A. Gunst et al., Astronomy&Astrophysics, vol. 556, A2.

[15] Dewdney P., (2013) SKA1 System Baseline Design SKA Project Documents.

[16] Dewdney P., P. Hall, R. Schilizzi, and T. Lazio, (2009) The Square Kilometre Array, Proc. of the IEEE, vol. 97, no.8, pp. 1482–1496.

[17] Wise M., Alexov A., Folk M. et al., (2011) Towards HDF5:Encapsulation of Large and/or Complex Astronomical Data, Astronomical Data Analysis Software and Systems XX., ASP Conference Proceedings, vol. 442, pp. 663.

[18] Fukugita M, Ichikawa T., Gunn J E, et al., (1996) Sloan Digital Sky Survey Photometric System. Astronomical Journal v.111, pp.1748, April.

[19] Ivezic Ž., Lupton RH., et al., (2004) SDSS data management and photometric quality assessment. Astronomische Nachrichten, vol. 325(6-8), pp. 583–589.

[20] Tody D.. (1986) The IRAF Data Reduction and Analysis System. In Lawrence D Barr, editor, 1986 Astronomy Conferences, pages 733–748. SPIE, August.

[21] Scott D, Pierfederici F, et al., (2007) The NOAO High-Performance Pipeline System: Architecture Overview. Astron. Data Analysis Soft. and Systems XIV, 376:265, October.

[22] Bilbao L., Lundin L., Ballester P., et al., (2010) Multi-Threading for ESO Pipelines. Astronomical Data Analysis Software and Systems XIV, pp.434:241, Dec,.

[23] Ranade A., Mahabalarao S., Kale S., (2007) A variation on SVD based image compression, Image and Vision Computing, vol. 25, pp. 771–777.

The Role of Cloud Services in Research Projects
(Case Study)

Tanya Dilkovska-Petrova[1]

Case Study: Anna Zagorska[2], Kristina Bliznakova[3]

[1]Haemimont AD, BIC-IZOT, Suite 425, Tsarigradsko Chaussee. 7th Kilometer, 1784 Sofia, Bulgaria

[2]Medical University, Sofia, [3]Technical University of Varna

[1]e-mail: tanya.dilkovska@haemimont.com

[2]e-mail: zagorska.anna@gmail.com

Abstract: Nowadays there are different ways to fund the researchers, mainly by research project grants. Proposals require good planning on two directions - financial and scientific. Often the performance of activities is done long before their approval. Then research projects are quite often compromised by circumstances that are not planned or under the control of researchers like equipment breakdown, staff leave, poor data quality and lost data, unexpected results, etc. Instead of concentrating on the tasks leading to the achievement of the project goals and the outcomes expected by stakeholders, the researchers need to handle blocking issue that may even fail the project. Although nowadays management plans for research projects include risks management strategies and prevention, issues are still occurring, especially in PhD projects. PhD students have a relatively short period of time to develop their research skills and produce a significant piece of work. Additionally the nature of research projects is such that the project will evolve significantly as the research progresses. It is even possible to get unanticipated results at some stage. This may introduce additional requirements and effort that is not included in the initial plan and may be in conflict with budget and timelines. This presentation will propose solutions to some common issues. It will first give explanation of cloud computing and services. Second it will show the benefits that cloud services can provide to researchers. Finally, a real case study will be presented in the area of clinical based research which output is targeted to outcome changes in clinical practice.

Keywords: cloud computing and services, research, project planning and management, success factors, Monte Carlo simulations

1 Introduction

Although cloud computing has become quite popular, people are still having different understanding what it is and how to best use it. The definition published by National Institute of Standards and Technology (NIST) outlines the most important aspects of cloud computing and can be used as a baseline for comparisons of cloud services and deployment models. A number of articles can be found that discuss the advantages offered by the different types of cloud services and how business can leverage them. This publication will prove that not only large companies can increase their profit and optimize their performance. Cloud services give new opportunities to small and middle businesses, start-ups, research projects and especially PhD students. It will be shown how some of the most common issues faced by researchers can be addressed with the help of cloud services. One true example will be used for illustration.

2 Cloud Computing Characteristics

Cloud computing, it is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model consists of five essential characteristics, three service models, and four deployment models. The collection of hardware and software that enables the five essential characteristics of cloud computing is called a cloud infrastructure. The cloud infrastructure can be viewed as containing both a physical layer and an abstraction layer. The physical layer consists of the hardware resources that are necessary to support the cloud services being provided, and typically includes server, storage and network components. The abstraction layer consists of the software deployed across the physical layer, which manifests the essential cloud characteristics. Conceptually the abstraction layer sits above the physical layer. [1]

Essential Characteristics

On-demand self-service. A consumer can provision computing capabilities as needed automatically without requiring human interaction with each service provider.

Broad network access. Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

Resource pooling. The provider’s computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.

Rapid elasticity. Capabilities can be elastically provisioned and released to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

Measured service. Cloud systems automatically control and optimize resource use by leveraging a metering capability, typically done on a pay-per-use or charge-per-use basis, at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

Service Models

Software as a Service (SaaS). The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

Platform as a Service (PaaS). The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

Infrastructure as a Service (IaaS). The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls).

Deployment Models

Private cloud. The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

Community cloud. The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.

Public cloud. The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.

Hybrid cloud. The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

3 Cloud Computing Advantages

Cloud services are becoming more and more popular during the past few years since they have first appeared on the IT market. The reason is the list of advantages they offer both to end users and businesses of all sizes. The obvious huge advantage is that you no more have to support the infrastructure or have the knowledge necessary to develop and maintain the infrastructure, development environment or application, as were things up until recently. The burden has been lifted and someone else is taking care of all that. Businesses are now able to focus on their core business by outsourcing all the hassle of IT infrastructure.

Let’s discuss some of the important advantages of cloud computing in more detail.

Cost efficiency. This is perhaps the most significant cloud computing benefit. Businesses and organizations of any type and size or even individuals can lower their IT expenses by leveraging the typical scalable and convenient charging models such as one-time-payment and pay-as-you-go. Consumers can save on licensing fees and at the same time eliminate overhead charges such as the cost of data storage, software updates, management etc. The lack of on-premises infrastructure also removes their associated operational costs in the form of power, air conditioning and administration costs. Consumers pay for what is used and disengage whenever they like. There is no need for capital expenditure at all. And because cloud computing is much faster to deploy, businesses have minimal project start-up costs and predictable ongoing operating expenses. Cloud computing allows you to forget about technology and focus on your key business activities and objectives. It can also help you to reduce the time needed to market newer applications and services.

Backup and recovery. When companies start relying on cloud-based services, they no longer need complex disaster recovery plans. With a managed service platform, cloud computing is much more reliable and consistent than in-house IT infrastructure. A cloud deployment is usually built on a robust architecture thus providing resiliency and redundancy to its users. Most providers offer a Service Level Agreement which guarantees 24/7/365 and 99.99% availability. Consumers can benefit from a massive pool of redundant IT resources, as well as quick failover mechanism - if a server fails, hosted applications and services can easily be transited to any of the available servers.

Cloud computing providers take care of most issues, and they do it faster. Aberdeen Group found that businesses which used the cloud were able to resolve issues in an average of 2.1 hours, nearly four times faster than businesses that didn't use the cloud (8 hours). The same study found that mid-sized businesses had the best recovery times of all, taking almost half the time of larger companies to recover.

Some 800,000 laptops are lost each year in airports alone. This can have some serious monetary implications, but when everything is stored in the cloud, data can still be accessed.

Scalability and performance. Scalability is a built-in feature for cloud deployments. Cloud is the flexible facility that can be turned up, down or off depending upon circumstances. Hand in hand, also comes elasticity, since cloud resources can be scaled to meet your changing IT system demands. This has been promoted as one of the important reasons to move to the cloud.

Regarding performance, the systems utilize distributed architectures which offer excellent speed of computations. Again, it is the provider's responsibility to ensure that your services run on cutting edge machinery. Instances can be added instantly for improved performance and customers have access to the total resources of the cloud's core hardware via their dashboards.

The cloud can accommodate and store much more data compared to a personal computer and in a way offers almost unlimited storage capacity. It eliminates worries about running out of storage space and at the same time it spares businesses the need to upgrade their computer hardware, further reducing the overall IT cost.

Collaboration efficiency. Public clouds offer services that are available wherever the end user might be located. This approach accommodates the needs of users in different time zones and geographic locations. Cloud applications improve collaboration by allowing dispersed groups of people to meet virtually and easily share information in real time and via shared storage. Employees can even chat to each other whilst making changes. This capability can reduce time-to-market and improve product development and customer service. Cloud computing is in that way especially appealing to international companies. If a company doesn't use the cloud, workers have to send files back and forth over email, meaning only one person can work on a file at a time and the same document has tones of names and formats.

A survey by Frost & Sullivan found that companies which invested in collaboration technology had a 400% return on investment.

According to one study, "73% of knowledge workers collaborate with people in different time zones and regions at least monthly".

Flexibility of work practices and easy adoption. Cloud computing services can be accessed via a variety of electronic devices that are able to have access to the internet. These devices include not only the traditional PCs, but also smartphones, tablets, laptops etc. Additionally as long as employees have internet access, they can work from anywhere. This flexibility positively affects knowledge workers' work-life balance and productivity.

A cloud system can be up and running in a very short period, making quick deployment a key benefit. On the same aspect, the introduction of a new user in the system happens instantaneously, eliminating waiting periods.

Furthermore, software integration occurs automatically and organically in cloud installations. A business is allowed to choose the services and applications that best suit their preferences, while there is minimum effort in customizing and integrating those applications.

Cloud applications usually entail smaller learning curves since people are quietly used to them. Users find it easier to adopt them and come up to speed much faster. Main examples of this are applications like GMail and Google Docs.

Cloud computing provides enhanced and simplified IT management and maintenance capabilities through central administration of resources, vendor managed infrastructure and SLA backed agreements. IT infrastructure updates and maintenance are eliminated, as all resources are maintained by the service provider. You enjoy a simple web-based user interface for accessing software, applications and services – without the need for installation - and an SLA ensures the timely and guaranteed delivery, management and maintenance of your IT services.

Access to automatic updates. With SaaS, the latest versions of the applications needed to run the business are made available to all customers as soon as they're released. Immediate upgrades put new features and functionality into workers' hands to make them more productive. What’s more, software enhancements are typically released quite frequently. This is in contrast to home grown or purchased software that might have major new releases only once a year or so and take significant time to roll out.

In 2010, UK companies spent 18 working days per month managing on-site security alone. But cloud computing suppliers do the server maintenance – including security updates –themselves, freeing up their customers’ time and resources for other tasks.

Environmentally friendly. With fewer data centers worldwide and more efficient operations, we are collectively having less of an impact on the environment. Companies who use shared resources improve their ‘green’ credentials as they are decreasing their energy consumption and carbon emissions with at least 30%, as for small companies even up to 90%, compared to using on-site servers.

4 Research Project

There are various factors that can bring a research project to success or failure. The ability of researchers to plan, coordinate and perform the research is a key point although some projects may succeed even when they are not formally managed. But usually an unplanned approach leads to stress in research team members, crises management when deadlines are not effectively met and the lack of time to deliver valuable outcomes.

This section will outline the primary items that give the recipe for a successful project, the most common issues faced during the execution of a project plan and how they can be overcome with the adoption of cloud services.

Success Factors

Clear goals. Before you begin any project you must have a clear understanding what you are trying to achieve, what your main goals are. An important aspect is to distinguish between outputs and outcomes. Outputs are the deliverables of the project, while outcomes are what happens as a result of the outputs. Once the desired outcomes are defined, the needed outputs can be determined from them. Having started with the end in mind your research efforts could be more focused.

Good management plan. Although not all grant proposals require a management plan, typically it is desired for large ones and can be beneficial for the rest. A good management plan should include at least the following:

- An outline of the project’s objectives and goals
- A list of actions to achieve the goals and produce the desired outcomes

- Descriptions of the roles and time commitments of the participants, as well as how these roles might change throughout the project
- Procedures to acquire and maintain equipment
- A timeline for various stages of the project
- A process to handle possible project modifications
- Consideration of the project's broader impacts

Stakeholders planning and management. A stakeholder is somebody who has an interest and involvement in the project. Poor stakeholders management may lead to communication issues and even cancellation of the project. First you need to identify the stakeholders and separate them into groups depending on their interest and power over the project, their ability to influence the desired outcomes and how they are achieved. Having completed this, you have identified who you need to monitor, to keep informed, to keep satisfied and to manage closely to prevent issues.

Timeline. A timeline can provide a sense of the proposed length of the stages of a project. It will let you establish intermediate deadlines and identify parallelization options. But perhaps the most important aspect is to determine the critical path activities which will have a negative impact on both the project and the researcher's career.

Acquisition and maintenance of equipment. Each project requires certain equipment and instrumentation for its execution. First you need to identify it, determine how it will be acquired and who will be responsible for its acquisition and maintenance. Next you must define who will have access to it and establish a plan to train the people that will be working with the equipment requiring special qualification. Finally, you must prepare steps to garantee availability of the equipment and find a place where it will be housed.

Risk management. Every project faces risks that might impact its success. The secret to managing risk well is to have thought through what might go wrong and prepare strategies to handle the key risks. First identify all the risks you can think of. Next, measure each one's impact based on its nature, costs and schedule impacts and decide on its likelihood to happen using a scale. Finally, plan how to manage the most likely and the most affecting risks.

Budget management. Effective budget management includes several aspects: understanding the commitments, understanding which part of the budget is maneged by who and keeping track of the spend. At any moment you must have clear picture on what is committed to be spent in the future, what have already been spent and how much is left to play with.

Limitations & Issues Addressed by Cloud Services

Following are the most common issues that even a great management project plan cannot fully eliminate but you can easily address them with the help of cloud services.

Longer task execution time. Underestimated task execution time is one of the most common issues especially in research projects which are unpredictable and evolving in many aspects by nature. The elasticity of cloud services though will let you play with resources configuration until you figure out what you need to achieve the desired performance.

Unexpected results. Unanticipated results due to mistake, both during execution, or during planning and analyses, may significantly break the timeline and change the initial plan. To handle this situation you will either need to find and eliminate the reason and rerun the tasks, or perform new additional tasks to understand what you have missed and how things should be changed. In both cases to stay on schedule you need a quick and easy way to change the capacity of the equipment you depend on for a short period until the issue is solved. Cloud services will not only provide you huge resource pool, rapid elasticity, ease of use and unlimited access from anywhere, but you will pay for it as long as you need it.

Poor data quality. Data quality may become poor if data is not well collected, organized or stored. To prevent this you can use Backup-as-a-Service, Storage-as-a-Service or Software-as-a-Service applications specifically for your needs.

Staff leave. Losing a member of the research team may result in several other issues: new tasks may need to be assigned to other team members, new person with specific qualification needs to be found, trained and included in the project, and data and know-how may be lost. This risk can be managed by using reporting system regularly, training of other project staff in techniques so no knowledge is kept by only one person, use of additional computational resources in the cloud so more tasks got parallelized and executed by less people.

Equipment breakdown. This is no more your problem if you are using cloud infrastructure as it is maintained by the provider. To guarantee local devices, for example laptops, you can use Backup-as-a-Service to make sure you have a copy of your recent local data in the cloud that can easily be restored on another device.

Underestimated budget. Cloud services are providing pay-as-you-go and pay-per-use charging models and full reporting at any moment for all liabilities that gives ability to effectively control, manage and predict your expenses.

Beaten to publish. The optimized performance of the distributed architecture, the elasticity, the guaranteed availability and security of cloud services are increasing your concurrency and significantly decreasing this risk. All you need to do is just to keep an eye on your competitors and use the benefits of cloud computing as much as possible.

5 CASE STUDY: Towards research solutions using Monte Carlo simulations - Cloud Services in favour of PhD Research Project

To design an experiment, plan or model that defines a problem, to test potential resolutions and to implement a solution is a part of work on a PhD thesis. The aim of this manuscript is to describe a part of a real research project with its problems and their solving.

Research experiment design

Cataract is the loss of transparency of the eye lens. It is not only age related, but associated with a prerequisite like ionizing radiation. In the focus of interest are medical specialists – radiologists, whose working place is within interventional room at close proximity to patient and X-Ray unit performing prolonged fluoroscopically guided procedures [3]. Moreover, some authors have shown that such procedures are often conducted without the use of the appropriate radiation protective equipment or with an inappropriate use of the latter [4,5]. On the other hand in recent years there has been an increased practice of interventional radiology and that results on increased doses at unshielded parts of the body and especially on eye lens.

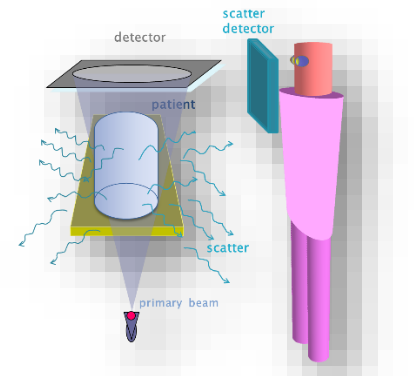


Figure 1. Schematic presentation of the setup used in the simulation study.

To estimate the dose distribution in the eye of a medical specialist in the Department of Medical Physics and Biophysics at Medical University, Sofia a research project was started. The first step of this study was to calculate the scattered energy spectra distribution at the level of the operator's head and distribution within the eye lens. However, the study of the dose distribution in the eye is only possible with MC simulations. For this purpose, an in-house developed Monte Carlo-based computer application was used to design computational phantoms (patient and operator), the acquisition geometry as well as to simulate the photon transport through the designed system (Fig. 1) [6].

For the purposes of the study, computational phantoms of a patient and operator were presented by simple geometrical primitives: cubes, cylinders, ellipsoids. Than the simulated system includes more than 20 objects built with composition of water, eye lens, aluminum, etc. To calculate the following detailed information was considered: photon energy, position and direction. All data processing and visualizations were realized in Matlab. The initial incident spectra from X Ray unit were calculated for 90 kV tube voltage, 12° angle tungsten target, 3 mm Al inherent filtration with an additional copper filtration (Cu) (0.1 mm, 0.3 mm, 0.6 mm, 0.9 mm) according to IPEM 78 [7].

A comparison beteen primary spectra for 90 kVp vs. scattered spectra is shown on Figure 2a,b. It shows the calculated scattered spectra, obtained from simulation study in comparison with the primary spectra for 90 kVp of X Ray unit with different filtration. A Matlab processing allows to calculate mean, effective and maximum energy.

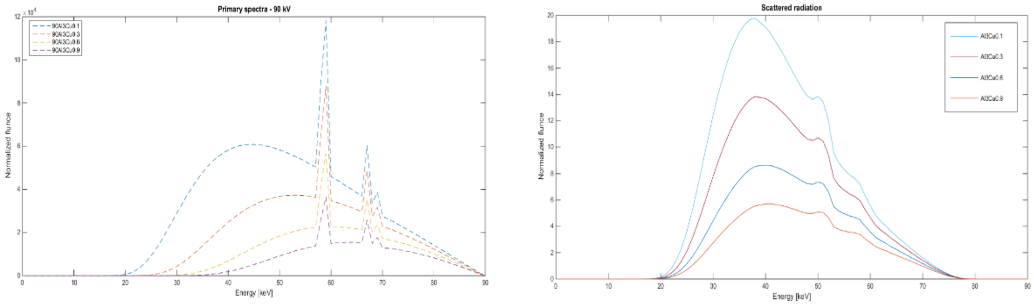


Figure 2. Comparison between primary beam and scattered radiation: a) primary beam spectra and b) simulated scattered spectra for 90 kV and different copper filters.

Cloud experience

As typically happens, working on a project is accompanied with different problems. Although there was a good collaboration between partners it wasn't possible to predict the computational time for all planned research tasks. Because of the huge volume of operated data while doing simulations, it was not possible to process and visualize it on a working PC. There were two possible outcomes: to report half work, which is not acceptable for any researcher, or to find a way out of the situation. To do simulations faster it was necessary to have more resources, but in the project was not predicted more funding for PCs. As a result we found cloud services as the only possibility.

On Table 1 is shown a comparison between used personal and cloud resources.

Table 1. Resources comparison

	Cloud Infrastructure	**Personal Computer**
Configuration	CPU: 8; RAM: 8GB; HDD: 70 GB	CPU: 1; RAM: 6GB; HDD: 1 TB
Number of runs in parallel	7	3
CPU % Usage	13	25
Memory % Usage	25	70
Environment availability hours	24/24	8/24

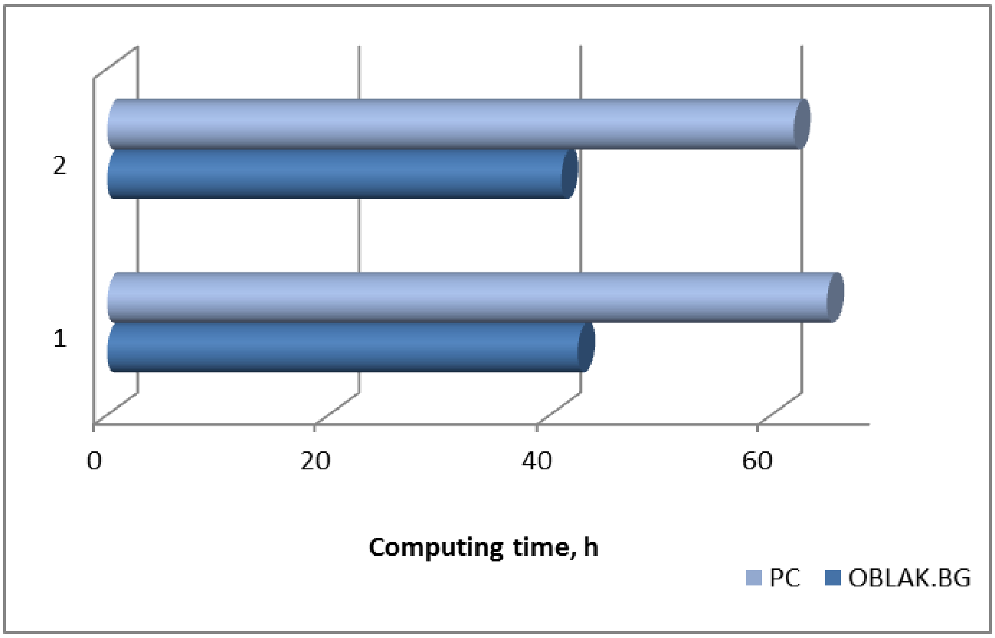


Figure 3. Computing time comparison:
1) 60 kV, 3 mm aluminum, 0.1 mm copper filtration, 57577 photons per run;
2) 70 kV, 4.5 mm aluminum, 0.3 mm copper filtration, 48 000 photons per run.

The cloud resources were chosen according to analysis specifics to obtain faster simulation process. To find the optimum resources configuration the processing time was recorded. To achieve the desired performance we needed to use more RAM and CPUs compared to the available local PC. The computing time was actually dependent on the number of photons to interact with the matter and the energy of the primary X-Ray spectra. The comparison of computing time for two scenarios is shown at Figure 3.The analyses showed more than 30 % time saving when cloud computing was used.

The cloud system can be accessed and managed from anywhere and at any time, which results in significantly shorter computing time and increased working process more than twice. At the same time working PC is free for processing and visualizing data. Combining these two types of resources optimization and parallelization were achieved.

6 Conclusions

In a conclusion the practice shows that cloud services are easy to use even for a non IT specialist, flexible according to research requirements and save time due to their broad access and full availability 24 hours a day, 7 days a week. At last but not at least they allow to pursue goals and to work over new once for a successful research.

7 Acknowledgements

This work was granted by Medical University – Sofia under Project 12D/2014, Contract 23D/2014. The authors of the case study express sincere gratitude to cloud services by Oblak.bg for granted infrastructure to perform simulations.

References

[1] Mell P., Grance T. (2011) The NIST Definition of Cloud Computing. *National Institute of Standards and Technology,* U.S. Department of Commerce, Special Publication 800-145.

[2] Henderson R. (2010) Research Project Management - Key Concepts, *My Consultants Ltd*.

[3] International Atomic Energy Agency TECDOC 1731. (2013) Implications for occupational radiation protection of the new dose limit for the lens of the eye.

[4] Vano E., González L., Guibelalde E., Fernández J. M. and Ten J. I. (1998b) Radiation exposure to medical staff in interventional and cardiac radiology, 71, Br. J. Radiol., 954–60

[5] Koukorava C., Carinou E., Ferrari P., Krim S. and Struelens L. (2011a) Study of the parameters affecting operator doses in interventional radiology using Monte Carlo simulations, 46, Radiat. Meas., 1216–22

[6] Cranley K. et al, 1997 IPEM Report 78: Catalogue of Diagnostic X-ray Spectra and Other Data (cd-rom ed.) The Institute of Physics and Engineering in Medicine (IPEM), York, UK

[7] Zagorska A., Bliznakova K., Buchakliev Z., (2015) Towards the estimation of the scattered energy spectra reaching the head of the medical staff during interventional radiology: A Monte Carlo simulation study, 2015 J. Phys.: Conf. Ser. 637 012036, http://iopscience.iop.org/1742-6596/637/1/012036

Big Data for Personalization – Opportunities, Applications

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: Big Data is the new trend in personalization. It arises as the latest focus in science, economics and society. It drives deep shift in the way that companies work, transforming the business organisation and management into data-centric. Therefore, all the institutions direct their efforts towards serving these demands. This research provides a brief view of Big Data driven opportunities for personalization and their applications in many diverse domains – science and education, healthcare, business and industry by presenting typical issues in collecting, processing, storing, and analyzing data.

Keywords: Big Data, Personalized services, Analytic tools

1 Introduction

The Big Data issue emerges from the exponential growth and wide availability of digital data that are difficult to be managed and analyzed using traditional tools and technologies. Recently it arises as a new focus in science, economics and society. It drives technology shift to data-centric structure and management in all areas. According to commonly used definitions the term Big Data relates to the extremely large amounts of diverse types of data usually collected through many devices and technologies such as smart cards, Wi-Fi sensors, RFID tags, etc., internet and social media (mainly raw, unstructured data – not complied with a specific pre-defined data model) or during routine actions such as statistics and accounting (structured data). Therefore, as Big Data ushers the society in an era of personalization and customization, all the organisations and institutions direct their efforts towards serving these needs and demands.

In order to unlock implicit potential of Big Data, a value of information in structured and unstructured data of huge volume and great diversity has to be discovered. This is tough task that involves the interdisciplinary methods for advanced technologies development including computational power, machine learning, data analytics and decision making techniques. This work provides a brief view of Big Data driven opportunities for personalization and their applications in many society and economic domains by presenting typical issues in collecting, processing, storing, and analyzing data.

2 Big Data Sources and Analytic Tools

Recently more data and more types of data are collected and analysed. Because of the rapid ICT penetration into society the number of printed documents decline, but the electronic documents keep growing, along with image and voice fragments/ files. All the information is gathered from almost everywhere:

- paper and digital text documents, books and magazines, scanned archives;
- public media – images, voice, video, stream;
- e-commerce sites, POS systems;
- business applications – ERP, CRM and BI systems;
- social networks, public web and databases;
- smart phones, mobile apps
- sensor data – biometrical, geospatial and industry devices, household appliances.

Most of the companies publish large amounts of information into the public domain that make large volumes of data accessible to everyone. Generally, data come from business transactions, log registers, emails, social media, geospatial sensors, RFID, etc. Considering the value of this information, business can understand users' behaviour in context so that to predict correctly indicators and to qualify whether something works or not and why, in order to make reasonable decisions. Further business strategies adaptations and modifications can be done to increase market share and revenue. Interventions for user experience improvements can be undertaken to provide personalized services and boost customer loyalty.

On Figure 1 a general scheme of the Big Data operating is presented. The information is collected from all possible sources: sensors, Internet, documents, sites, electronic media etc. It has to be classified, filtered, associated (pre-processed) through appropriate tools and prepared for further more sophisticated processing comprising analytical functions as pattern and structure discovering, outlier detection (core processing). Part of the collected and processed data is utilised for common purposes (statistical researches, creating policies) and another part is used to personalize delivered services. Data can be processed in real time or not depending on the actual application. Cloud structures are successfully employed when implementing Big Data. Data as well as tools can be situated in one or more private and public clouds of different providers, thus reducing costs.

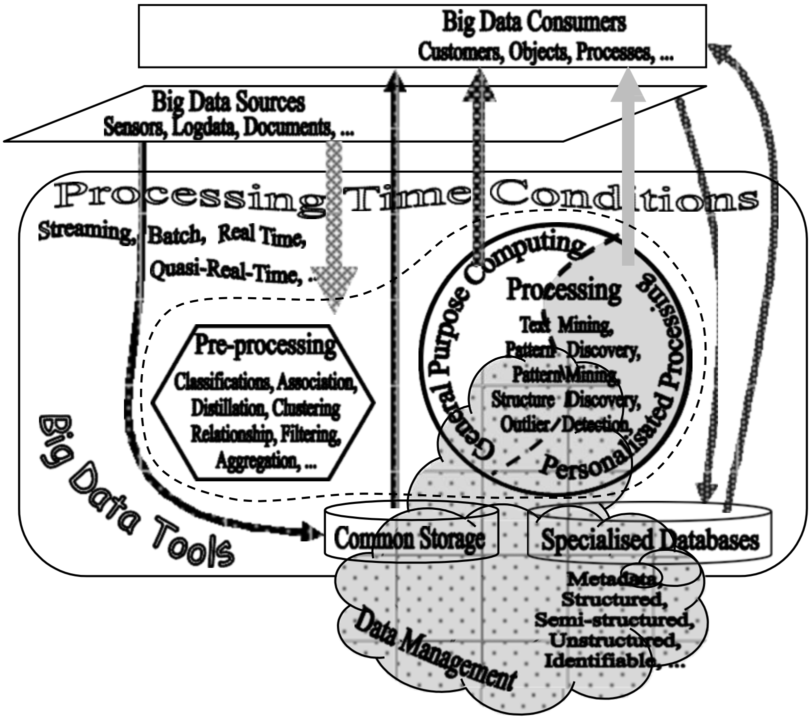

Big Data Consumers
Customers, Objects, Processes, ...
Big Data Sources
Sensors, Logdata, Documents, ...
Processing Time Conditions
Streaming, Batch, Real Time,
Quasi-Real-Time, ...
Pre-processing
Classifications, Association, Distillation, Clustering Relationship, Filtering, Aggregation, ...
General Purpose Computing
Processing
Text Mining, Pattern Discovery, Pattern Mining, Structure Discovery, Outlier Detection,
Personalisated Processing
Big Data Tools
Common Storage
Specialised Databases
Metadata, Structured, Semi-structured, Unstructured, Identifiable, ...
Data Management

Figure 1. A concept of general functional model of Big Data

The Big Data analysis tools are becoming increasingly important to competitiveness for the subjects in all industry areas as their adoption continues to grow. In order to reveal their full potential this Big Data tools have to [2, 5]:

- allow getting quick answers across large data sets;
- allow integration with existing tools and systems (Databases, ERP, CRM, etc);
- provide better security options trough access management;
- conform with standards and best practices for developing and using;
- provide industry with solutions that utilize Big Data everywhere;
- define a wide variety of use cases, suitable for all kinds of enterprises;
- provide easy readable and comprehensive information for problems solving, improved analysis and more accurate predictions;
- be scalable to address different needs;
- be easy to use (not requiring programming skills) and give conscious notifications.

One the most commonly used tools for storage and processing of large sets of data is Apache Hadoop [7]. It enables businesses to acquire deep insight from immense quantity of structured and unstructured data. The work over Big Data is performed in distributed manner across clusters of computers using various programming models. The data are analysed by the number of powerful Hadoop’s ecosystem components such as Cloudera Impala query engine, MapReduce programming model, and others from the collection of additional software packages. The result is fast, high-performance, cost-effective data management for real-time business applications.

Another data processing tool is offered by Google as interactive analytical service BigQuery and cloud service Google Cloud Storage [6]. The company converts the data collection and analysis into business model – despite it provides these services for free, at the same time it takes benefit from the collected information, while watching how these services are being used.

3 Applications

Companies which scope of business is data driven connected will have great benefit from Big Data implementation. They can take advantage from industry-specific analytics that will help them to reduce time and costs of transaction processing. Some of the most common applications of Big Data analytic tools are described below [1, 2, 4, 5, 8]:

- *Log analysis* of huge bulk of logs and traces in IT systems that is impossible even to read, now can be thoroughly examined and analysed so to be at the core of evidence-based decision making. The benefits will be in two main directions: diagnosing, identifying and preventing problems and optimizing the performance.
- *Real time analysis* allows performance evaluation and information assessment to be done along with analysis of the whole business environment data – already stored and ongoing data. These tools are widespread in financial and stock exchange, intelligent transport and logistics traffic management, security and crowd control, emergency and disasters warning, etc.
- *Predictive analytics* of historical transaction data in retail chains, tourist and utility services, etc. provides valuable information about type and quantity of sought goods/ services, their availability, most profitable products, etc. The Big Data tools allow for automatic demand forecasting, planning of supply and advertisement, etc.
- *Customer relationship management* is particularly important for detection of repeating and persisting problems or behaviour patterns in real time not only by making sense of time/ quality, but also by considering the context as a whole. This is a proper indicator of real company market place, but without a Big Data analytics, much of these insights will be unseen or recognized too late.

- *Fraud detection and risk management* in insurance and commerce business is one of the most compelling problems. Big Data tools allow real time analysis of claims and transactions and recognize repeated patterns and behaviour within transaction database so that to discover abnormal events.
- *Social media activity analysis* provides real-time market products perceptions. Big Data tools can gather and evaluate customers' insights, so that companies can ad hoc optimise their policies concerning prices, advertisements and promotions.

3.1 A Scheme of Big Data Flow

Figure 2 illustrates the information flow in a generalized system using Big Data. It is apparent that in the actually considered area the data collected can be processed and used for the creation of both – strategies for an entire area policies and individual approach to every user of the products. In the former case the operations pertain to the macro-level and in the second - to the micro-level (personalization) [11].

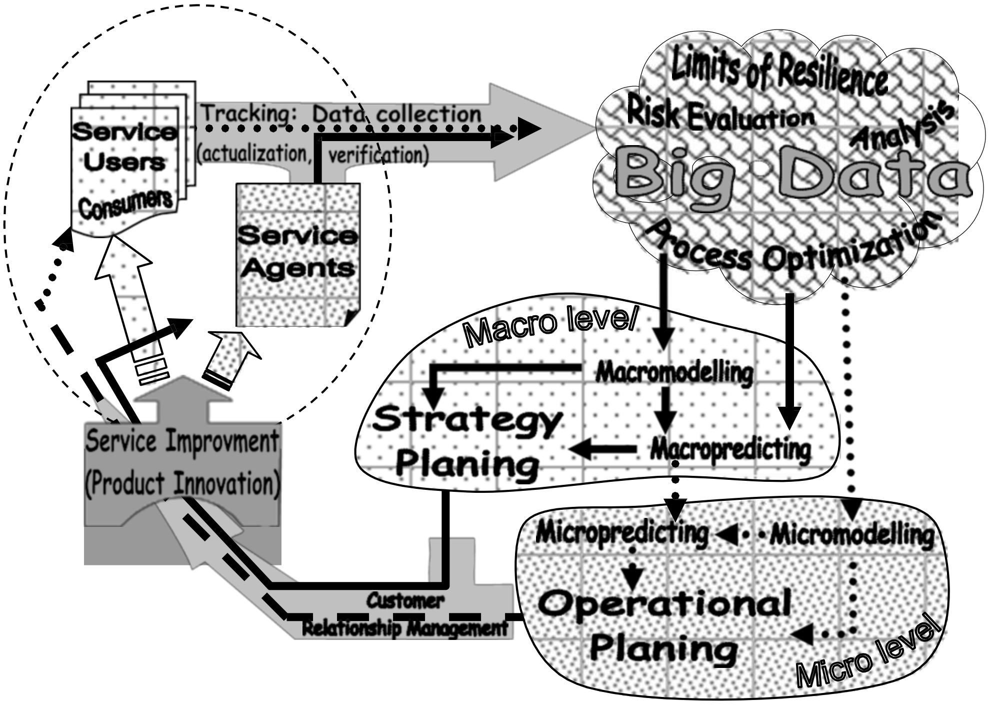


Figure 2 Big Data information cycle

Data are gathered continuously (in real and quasi-real time, or otherwise, e. g. historically) from each product user or service agent acting in the area. Users (consumers) can be students (in the educational area), patients (in the health care area), clients (in shops, banks), passengers (in transport), etc. The service agents (product providers) are respectively educators, medical staffs, sellers, traders, bankers, managers etc. Through various methods and tools inherent for Big Data, the collected data can be processed (summarized, classified, clustered) according to various indicators essential for the treated area (e.g. by residence, by profession, etc.) to build models, simulate situations, explore reactions and thus to make predictions about future processes, etc. These activities relate to the macro-level and are relevant to future developments in that particular area. On the other hand, data on personal characteristics and preferences of each individual enable personalized tackling, again using mathematical models and predictions. For example, in health, at a macro-level we can envisage potential diseases (related to environment, profession, etc.). At micro-level we may administer for each patient appropriate prevention or treatment based on his / her specific health status.

In the educational area at a macro level one can comply the scope of training material in certain disciplines according to the average knowledge level in a region (extend / shorten the volume, depth). At the micro level for the particular student we can select the appropriate teaching course to attain the maximum positive learning impact.

Data obtained from both the macro- and micro-level is used to improve, renovate, develop and create new product offerings in the particular area.

3.2 Education

The exponential growth of educational data during last decade is generated mainly from online sources during the interaction of students with e-learning platforms and reflects the individual learning process. This information represents "snapshots" of the students’ learning behaviour [3]. These are log- data accumulated from the actions in learning platforms concerning learning resources usage, interactions and cooperation between learners and teachers, participations in education forums, thematic groups and other social networks. The education professionals also generate data – they create, annotate and store learning resources in web-based repositories, build semantic networks and conceptual maps. On the other hand, the information in traditional education mainly comprises personal and administrative issues – biographical data; learning style and psychological type; training experience, qualifications and skills; academic degrees and majors; curriculum, programs and courses, etc. [11]. All these educational data are at a hand, but not always easily available, intelligible and comprehensible, which necessitates Big Data analytic tools for their utilisation. Extraction and processing of such large volume heterogeneous information is extremely complex task, so it employs mathematical methods and learning analytics.

To improve learning process the traces leaved behind by students’ actions are collected and used. The gathered data and their correspondent context about learners’ behaviour are measured and analyzed to understand and optimize the education. The procedure of obtaining an insight view through problem definition and application of statistical models and analysis is defined as “Learning Analytics”, which comprises web and academic analytics, educational data mining, etc. As a result, personalized learning resources and courses are provided and more efficient education is achieved.

In educational context according to Figure 2, data collected at macro level represents aggregated information concerning curricula, courses, learning resources, etc. It serves for evaluation of quality and efficiency, also for educational process analysis regarding strengths and weaknesses. The specific characteristics and trends are determined so that educational programs and policies to be improved.

A complete picture of individual learning path concerning learner’s behavioural data and perception style in the whole educational process is obtained at micro-level. They encompass actions performed in problems solving (approach, speed, trials, mistakes, decision quality, help, etc.). Further these data are analysed and comprehensive information on the personal learning characteristics of each student (strengths and problematic sides, specific needs, abilities, aptitudes, interests) are revealed. This is the basis for the learning process personalization and resources’ adaptation – selection of suitable learning resources, provision of optional educational materials, application of individualised teaching, and arrangement of customised curricula. In addition, the feedback ensures data-driven personalized decision-making about the whole educational process

3.3 Healthcare

By using a wide variety of methods for processing and analysis of structured, semi-structured and unstructured data the Big Data phenomenon is particularly suitable for application in healthcare [9]. The customization ability dramatically changes the work in this area. The objective function is to optimise health management, based on the plurality of collected data, starting from the pre-natal period. Especially in case of chronic illness, continuous data collection is crucial to provide personalized care for patients. Data are differently processed and subsequently analyzed and the results can be used for a prevention, prophylaxis, treatment, recovery, immunizations, etc., and to assist in decision making. The aggregation of personal data can be used for extraction of statistical information and discover community, local, age, family and other dependencies, trends, problems. In practice, the data sets can be used further as data with new quality and surplus value [9, 12, 13]. Big Data in health care unites internet consumption on health occasions, use of social media, use of stationary and mobile health applications – all data types and their derivatives are stored and can be extracted, statistically processed, analyzed and used in a new context. Availability of more authentic data allows achieving objective expert information that reflects positively on the quality of decision making.

The use of Big Data in health care requires compliance with all regulations, standards and rules for working with digital data. Strict control has to ensure the use of collected information solely for the purposes of health. The access must be regulated and controlled, especially for legally protected data.

The security management process is also extremely important. Ensuring imposes formal regulation at all levels – legal, regulatory, manufacturing, market regulation, professional, public. Informal regulation is needed too – moral, ethical; and last but not least implementation of international IT certifications, standardization and rules for virtual communication.

Patient data can be collected from clinical records, tissue banks, personal genetic data and genomic data sources, environmental and geospatial data sources, health maintenance apps [13]. The personal profile information comprises generic/ family specifics, blood type, Rh, vaccinations, examinations carried out, surgery performed, diseases, treatments, used drugs and reactions, vitamins and dietary supplements intake, diet, suffered accidents, personal habits, and motor skills.

Big Data tools empower decision making in personalized healthcare by augmenting such significant key factors as:

- history statements – medical and family – diseases, treatments, medical tests, vaccinations, accidents, surgery interventions, etc.;
- clinical genomic data – genetic status that is important for medical decision making, selection of supportive care, prognostic considerations, reproductive decisions, etc.;
- pattern of inheritance, affected age group;
- environmental factors – generally ignored in a traditional healthcare setting, but they influence the individual genotype formation and may play a key role in understanding the personalized medicine outcomes.

Through appropriately processed medical data a complete view of the patient health can be achieved. Applying Big Data analysis tools a personalized system can be developed for decision support. Furthermore this information can be used to identify similarities between medical cases and in this way facilitate and accelerate diagnostic and treatment.

The collected personal data can be used for statistical purposes to form an exact health image of a community, a population in a region, age related trends thus by relevant modelling to detect risk factors and processes in order to make evidence-based prognostics.

3.4 Business and Trade

In the era of globalization, managers already have tools to collect and share information on a frequent, real time basis, so business organisations have plenty of Big Data, but they also need tools for their interpretation and utilisation in order to make sensible decisions. It is hard work to transform such huge amount of data into useful information and there have to be experts to make analysis and forecasts. This challenge can be overcome with the help of modern natural language generation (NLG) and artificial intelligence that can present computer data in human-understandable reports, summaries or analysis [2, 4]

In fact, if it is possible to understand customer’s sentiment about a particular company and its products, it would be possible to predict future earnings and price moving, which will be a great advantage for investors [2, 4, 5]. Big data tools crawl through a various data sources with structured and unstructured information from purchase histories, customer relationship management, social media, credit cards, trade and travel, etc.

One of the biggest online retail companies - Amazon has already put Big Data into practice. The company has made huge investments in services based on commercial activity data. It provides seller consultant with information about each customer’s transactions. Besides, the stored data concerning related product details, recommendations, reviews and comments allows him to make immediately reasonable personalized further offers. The next step is to send latest books, DVDs, video games and devices before purchasing, based on previous customer orders.

Another example is Walmart chain of hypermarkets which keep a track of each customer’s purchases. They capture and store in real time this huge amount of data along with products description, store-location, sales count, etc. in data repositories. In that case Big Data analysis becomes indispensable for effective targeting both the existing and new customers to make more purchases from Walmart [10].

3.5 Internet of Things

The term Internet of things (IoT) was created by Kevin Ashton in 1999, to describe a system where the Internet is connected to the physical world via sensors. IoT is a vision of “ubiquitous computing” or the Internet of anything, anytime and anywhere, e.g. various “smart” objects, not only computers, can send and receive data trough remote access via the network [2]. Nowadays more and more appliances, equipments and systems have embedded smart functions that continuously generate data and influence social and economic life in the way activities are done, things are manufactured, sold, and even used.

Many companies embrace the IoT and develop technical devices equipped with technology to interact with each other and with the external environment without human intervention. The investments in so called industrial internet are in production not in consumer goods. The idea is to build a network of sensors and intelligent systems combining several units, power plants, vehicles, etc. [2]. The engines equipped with sensors networks can monitor the smallest changes and receive feedback in order to do automatically fine tuning, to increase efficiency, etc. The same concept is followed in the development of power supply units, medical equipment and domestic appliances. Internet of things and Big Data are closely related and frequently used together, so big international companies invest millions in using Big Data technologies. Smart devices enable computers to observe, identify and "understand the world" without the need of human intervention and has the capacity to transform the whole business.

4 Conclusions

Recently Big Data tools and analytics are rapidly becoming common in the economics and society. They enable fundamental shift to data driven management. Business and social institutions such as education and healthcare apply Big Data analytics to provide personalized services, measure the market response, optimize the management, etc. Wide range of companies is also starting to adopt such institution-level analyses in order to detect areas for improvement, set policies, and evaluate their effects. Measuring and making visible performed actions enable them to develop skills in intervention monitoring and to see directly the results.

Furthermore, managers gain a complete view on company's performance that helps to improve both the control and management. Big Data analysis also provides them the opportunity to see quickly the effectiveness of their actions and through feedback to reflect individual needs and requirements. Generally, if managers could understand the possibilities of Big Data tools, they could employ their power at the individual level. Moreover, mobile devices already allow easy access to cloud services, data mining and analytics tools which is another prerequisite to management to be always on.

References

[1] Amatriain X. (2012) Mining Large Streams of User Data for Personalized Recommendations. In: *SIGKDD Explorations,* 14, 2, 37-48.

[2] Bessis N., Dobre C. ed. (2014) *Big Data and Internet of Things: A Roadmap for Smart Environments,* Studies in Computational Intelligence Volume 546, Springer.

[3] Byun, J., Pennington D.; Cardenas J.; Dutta S.; Kirwan, J. (2014) Understanding Student Behaviours in Online Classroom – Data Scientific Approach. In: International Congress on Big Data, IEEE, 802-803.

[4] Chen H., Chiang R. H. L., Storey V. C. (2012) Business Intelligence and Analytics - from Big Data to Big Impact. In: *MIS Quarterly*, 36, 4, 1165-1188.

[5] Chen M., Mao S., Liu Y. (2014) Big Data: A Survey. In: *Mobile Networks and Applications*, 19, Springer 171–209.

[6] https://cloud.google.com/bigquery/

[7] https://hadoop.apache.org/

[8] Krumeich J., Werth D., Loos P., Jacobi S. (2014) Big Data Analytics for Predictive Manufacturing Control –A Case Study from Process Industry. International Congress on Big Data, IEEE, 530-537.

[9] Reddy Ch. K., Sun J. (2013) Big Data Analytics for Healthcare. In: SIAM International Conference on Data Mining, Austin, TX.

[10] Sharma S. Mangat V. (2015) Technology and Trends to Handle Big Data: Survey. In: International Conference on Advanced Computing & Communication Technologies, IEEE, 266 – 271.

[11] Terzieva, V., Todorova K., Kademova-Katzarova P. (2015) Big Data – Opportunities and Challenges for Education. In: BulDML ERIS2015-book, 136-145. available at: http://sci-gems.math.bas.bg/jspui/bitstream/10525/2440/1/ERIS2015-book-p14.pdf

[12] Viceconti M., Hunter P., Hose R. (2015) Big Data, Big Knowledge: Big Data for Personalized Healthcare. *Journal of Biomedical and Health Informatics*, 19, IEEE, 1209-1215.

[13] Yesha Y., V. P. Janeja, N. Rishe, Yesha Y. (2014) Personalized Decision Support System to Enhance Evidence Based Medicine through Big Data Analytics. In: I*nternational Conference on Healthcare Informatics* (ICHI), IEEE, 376.

Big Data in a Smart City Ecosystem: Models, Challenges and Trends

Rumen Nikolov[1], Boyan Jekov[1], Petya Mihaylova[2]
[1]UniBIT – Sofia, [2]Technical University of Sofia
r.nikolov@unibit.bg, b.jekov@unibit.bg, mihaylova_p@yahoo.com

Abstract: The Internet of Things (IoT) is a new paradigm that combines aspects and technologies coming from different approaches. Ubiquitous computing, pervasive computing, Internet Protocol, sensing technologies, communication technologies, and embedded devices are merged together in order to form a system where the real and digital worlds meet and are continuously in symbiotic interaction. By putting intelligence into everyday objects, they are turned into smart objects able not only to collect information from the environment and interact/control the physical world, but also to be interconnected, to each other, through Internet to exchange data and information. The expected huge number of interconnected devices and the significant amount of available data open new opportunities to create services that will bring tangible benefits to the society, environment, economy and individual citizens. This digitization of the urban space has led to a rich ecosystem of data producers and data consumers. Moreover, heterogeneous sources differ in terms of data complexity, spatio-temporal resolution and -curation /maintenance costs.

Keywords: big data, smart city, digital ecosystem, internet of things

1 Introduction

With the development of computer technology, there is a tremendous increase in the growth of data. Society is becoming increasingly more instrumented 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 key to competitive advantage. Analytics solution as data mining is the technique that can discover new patterns from large data sets. For many years it has been studied in all kinds of application area and thus many data mining methods have been developed and applied to practice. Big Data refers to various forms of large information sets that require special computational platforms in order to be analyzed. In summary, the design of big data systems keeps evolving when we need to handle larger-scale of data and more challenging user demands. Current cities are complex systems that are characterized by massive numbers of interconnected citizens, businesses, different modes of transport, communication networks, services and utilities. World is generating and using a big amount of data every single minute. The rapid growth faced by several cities has generated traffic congestion, pollution and increasing social inequality [9].

This paper is created based on the literature review of science articles published the rest three years. It represents models for big data managing and examinations.

The rest of the paper is organized as follow: Section 2 presents models and methods used in examination of both modern terms Big Data and Smart City. In Section 3 we are looking for the linkage between Big Data and Smart City and especially, how strong the linking is and what is its implication on both BD and SC. Section 4 represents two practical Study cases. In 5 are given trends and challenges of Big Data Management. Section 6 is conclusion.

2 Big Data and Smart City – Definitions and Models

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. There are several impacts of using the Big Data. For facing the competitions and strong growth of individual companies, it supports by providing them a huge potential. Certain aspects are needed to be followed so that we can get timely and productive results from Big Data because the precise use of Big Data can give the proliferation to throughput, modernization, and effectiveness for entire divisions and economies. To be able to extract the benefits of Big Data, it is crucial to know how to ensure intelligent use, management and re-use of Data Sources, including public government data, in and across country to build useful applications and services. It is crucial to evaluate the best approach to use for filtering and/or analyzing the data.

2.1 Big data Definitions

Big data definitions have evolved rapidly, which has raised some confusion. This is evident from an online survey of 154 C-suite global executives conducted by Harris Interactive on behalf of SAP in April 2012 (“Small and midsize companies look to make big gains with big data,” 2012). Fig. 2 shows how executives differed in their understanding of big data, where some definitions focused on what it is, while others tried to answer what it does. Clearly, size is the first characteristic that comes to mind considering the question “what is big data?” However, other characteristics of big data have emerged recently. For instance, Laney (2001) suggested that Volume, Variety, and Velocity (or the ThreeV’s) are the three dimensions of challenges in data management. The Three V’s have emerged as a common framework to describe big data (Chen, Chiang & Storey, 2012; Kwon, Lee & Shin, 2014). For example, Gartner, Inc. defines big data in similar terms:

“Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.”(“Gartner IT Glossary, n.d.”)

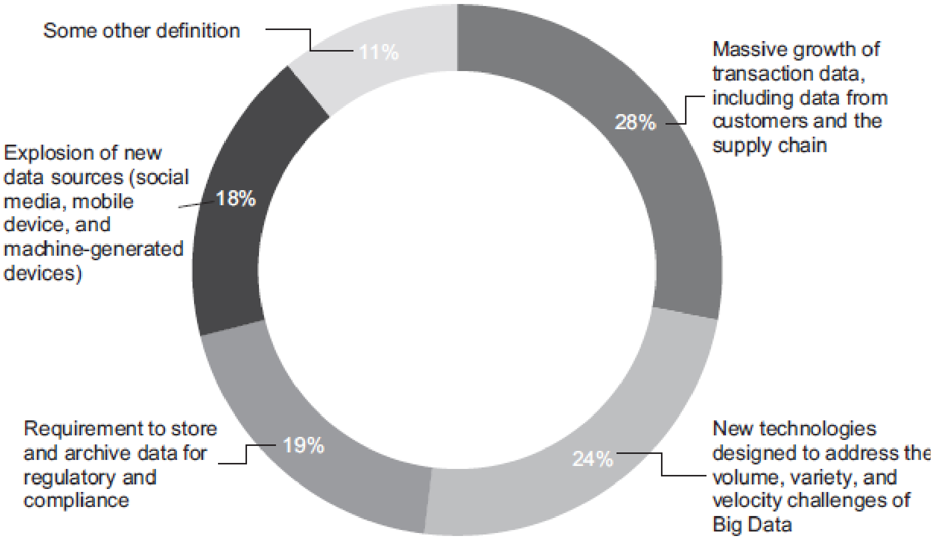


Fig.1 Definitions of big data based on online survey of 154 global executives in April 2012

Big Data is characterized by what is often referred to as a multi-V model, as depicted in Fig. 3. Variety represents the data types, velocity refers to the rate at which the data is produced and processed, and volume defines the amount of data. Veracity refers to how much the data can be trusted given the reliability of its source [1], whereas value correspond the monetary worth that a company can derive from employing Big Data computing. Although the choice of Vs used to explain Big Data is often arbitrary and varies across reports and articles on the Web – e.g. as of writing Viability is becoming a new V – variety, velocity, and volume are the items most commonly mentioned.

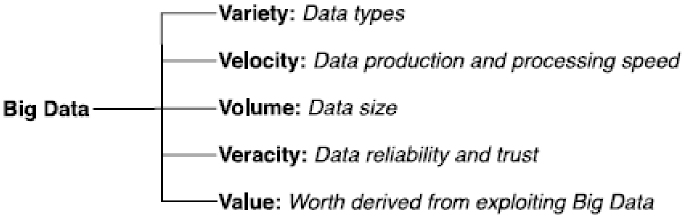

Big Data
Variety: Data types
Velocity: Data production and processing speed
Volume: Data size
Veracity: Data reliability and trust
Value: Worth derived from exploiting Big Data

Fig. 2 Some „Vs” of Big Data

Handling and analyzing this data poses several challenges as it can be of different types (Fig. 3). It is argued that a large part of data produced today is either unstructured or semi-structured. Considering data velocity, it is noticed that, to complicate matters further, data can arrive and require processing at different speeds, as illustrated in Fig. 4.

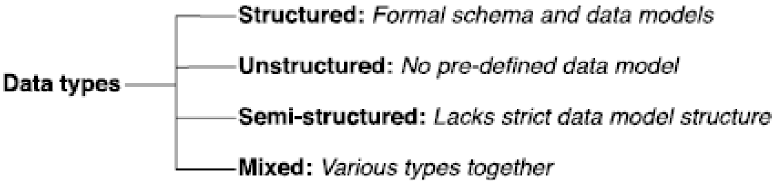

Data types
Structured: Formal schema and data models
Unstructured: No pre-defined data model
Semi-structured: Lacks strict data model structure
Mixed: Various types together

Fig. 3 Variety of Data

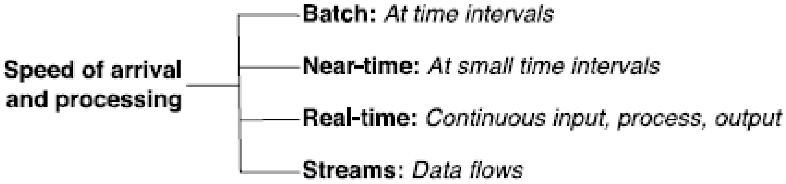

Speed of arrival and processing
Batch: At time intervals
Near-time: At small time intervals
Real-time: Continuous input, process, output
Streams: Data flows

Fig. 4 Velocity of Data

2.2 BD programming model MapReduce

Organizations need to build an investigative computing platform to realize the full value of big data. This enables business users to make use, structure and analyze big data to extract useful business information that is not easily discoverable in its actual original arrangement. The programmers use the programming model MapReduce to retrieve precious information from such big data. The main features and problems associated in handing different types of large data sets are summarized in the table below. It gives precise information how Big Data technologies can help solve them.

Table 1: Summarizes the main features, challenges and technology responses connected to handing different types of large data sets

Attribute	Features	Challenges and Skill responses
Volume	Amount of generated data has increased tremendously the past years. However, this is the less challenging aspect in practice.	Internet has created tremendous increment in the global data production. A response to this situation has been through the generalization of the cloud based solutions. The noSQL database approach is a response to store and query huge volumes of data heavily distributed.
Velocity	Production of data is growing with high speed and such produced data must be	Millions of connected devices (smartphones) are getting added daily which results in the increase of not only the volume but also velocity. To get a competitive edge, global companies

	collected in shorter time frames.	considered the Realtime data processing platforms as a requirement.
Variety	There came the explosion of data formats that range from structured information to free text with the multiplication of data sources.	The current way to collect and analyze non-structured or semistructured data is just opposite from the manner the traditional relational data model and query languages does. This reality has resulted in the evolution of new kinds of data stores that gives the ability to support flexible data models.
Value	Until recently, there was more focus on recording the large volumes of data but not bothered how to conquer them.	Big Data technologies are deeping their roots in creating, capturing and exploiting large volumes of data. In principle, the challenge comes while transforming underdone data into information that contains value and can be used in decision making or other business requirements.

MapReduce is designed to be used by programmers, rather than business users. It is a programming model, not a programming language. It has gained popularity for its easiness, efficiency and ability to control "Big Data" in a timely manner. The steps involved in working of MapReduce can be shown in as:

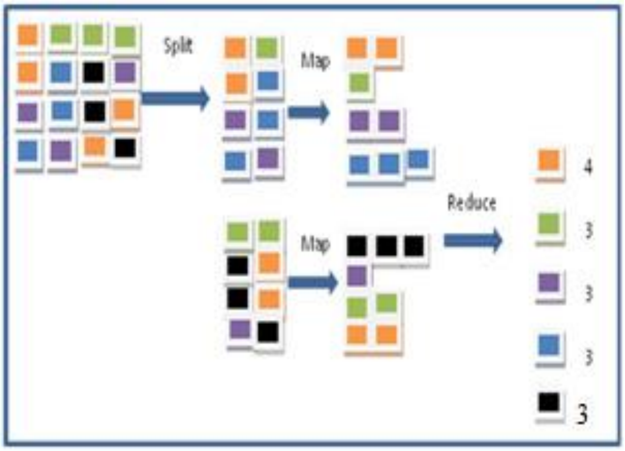


Fig. 5: Steps in MapReduce to process the database

The applications which include indexing and search, graph analysis, text analysis, machine learning, data transformation and many more, are not easy to implement by making the use of standard SQL which are employed by relational DBMSs. In such areas the procedural nature of MapReduce makes it easily understood by skilled programmers. It also has the advantage that developers do not have to be concerned with implementing parallel computing – this is handled transparently by the system. Although MapReduce is designed for programmers, nonprogrammers can exploit the value of prebuilt MapReduce applications and function libraries. The architecture of MapReduce can be depicted as:

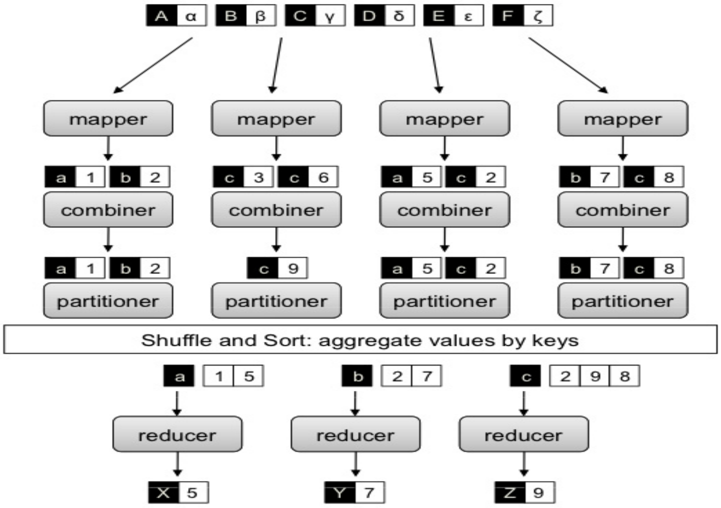


Fig. 6: MapReduce with combiners, partitioners

Table 2: Description of mappers, reducers, partitioners and combiners

Mappers	• Required to generate an arbitrary number of intermediate pairs
Reducers	• Applied to all intermediate values associated with the same intermediate key.
Partitioners	• Its main job is to divide the intermediate key space, and then to assign the intermediate key-value pairs to reducers.
Combiners	• Combiners are an (optional) optimization. • Before performing the phase of shuffle and sort, it allows the local aggregation of data. • Essentially, combiners are used to save bandwidth, e.g.: word count program.

MapReduce programs are usually written in Java. They can also be coded in other languages such as C++, Python, Ruby, R, etc. These programs may process data stored in different file and database systems. [2]

2.3 BD Analytics techniques

Analytics techniques allow the extraction of big data into meaningful insights. This efficient process is needed by the organizations because turns the high volumes of fast-moving and diverse data into meaningful result helping the process of decision-making. The overall process of extraction is divided in 5 stages, but roughly they can be aggregated into two basic groups – Data Management and Analytics. Data management involves processes and supporting technologies to acquire and store data and to prepare and retrieve it for analysis. Analytics, on the other hand, refers to techniques used to analyze and acquire intelligence from big data. Thus, big data analytics can be viewed as a sub-process in the overall process of 'insight extraction' from big data. [4]

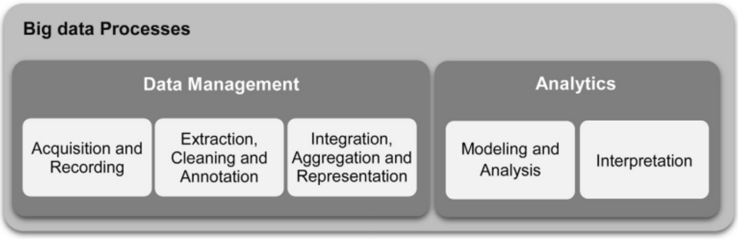

Big data Processes
Data Management
Analytics
Acquisition and Recording
Extraction, Cleaning and Annotation
Integration, Aggregation and Representation
Modeling and Analysis
Interpretation

Fig. 7. Processes for extracting insights from big data

Text analytics (text mining) refers to techniques that extract information from textual data. Text analytics involve statistical analysis, computational linguistics, and machine learning. Text analytics enable businesses to convert large volumes of human generated text into meaningful summaries, which support evidence-based decision-making. The used methods of Text Analytics are:

- ✓ Information extraction (IE) – extracts structured data from unstructured text
- ✓ Text summarization - automatically produce a succinct summary of a single or multiple documents
- ✓ Question answering (QA) - provides answers to questions posed in natural language
- ✓ Sentiment analysis (opinion mining) - analyze opinionated text, which contains people's opinions toward entities such as products, organizations, individuals, and events

Audio Analytics - Audio analytics analyze and extract information from unstructured audio data. Call centers use audio analytics for efficient analysis of thousands or even millions of hours of recorded calls. Audio Analytics uses two very common technological approaches:

- ✓ LVCSR systems (large-vocabulary continuous speech recognition) – it is automatic speech recognition and match sounds to words in predefined dictionaries.
- ✓ Phonetic-Based systems (phonemes) – Phonemes are the perceptually distinct units of sound in a specified language that distinguish one word from another (e.g., the phonemes/k/and/b/differentiate the meanings of “cat” and “bat”).

Video Analytics - Video analytics, also known as video content analysis (VCA), involves a variety of techniques to monitor, analyze, and extract meaningful information from video streams. According to the system architecture two approaches are known:

- ✓ Server-based architecture - the video captured through each camera is routed back to a centralized and dedicated server that performs the video analytics.
- ✓ Edge-based architecture - analytics are applied at the ‘edge’ of the system. That is, the video analytics is performed locally and enabling a more effective content analysis.

Social media analytics – analytics Social media analytics refer to the analysis of structured and unstructured data from social media channels

- ✓ Content-based analytics
- ✓ Structure-based analytics
- ✓ Community detection
- ✓ Social influence analysis

Predictive analytics – analytics Predictive analytics comprise a variety of techniques that predict future outcomes based on historical and current data

- ✓ Heterogeneity
- ✓ Noise accumulation
- ✓ Spurious correlation
- ✓ Incidental endogeneity

2.4 Data Management Cloud Models

One of the most time-consuming and labour-intensive tasks of analytics is preparation of data for analysis; a problem often exacerbated by Big Data as it stretches existing infrastructure to its limits. Performing analytics on large volumes of data requires efficient methods to store, filter, transform, and retrieve the data.

Management of analytics consists of three cloud models: private, public and hybrid. The private model is characterized with the highest level of control of security and data privacy. It is suitable for business sectors. Public model gives high efficiency and shared recourses with a low costs. It is perfect for the public sectors. The hybrid is combined model of the first two and provides recourses from the public sector to the private where is needed as assures the same high level of privacy and security of the used resource or service [1]. The three cloud models combined with the data provide four Data and analytic managements:

1) Data and models are private

2) Data is public, models are private

3) Data and models are public

4) Data is private, models are public

When the business models are managed and supported especially from providers, not only the services but the skills of data experts need to be managed. There are solutions of data management, focusing on storage and retrieval of data for analytics data diversity, velocity and integration; and resource scheduling for data processing tasks.

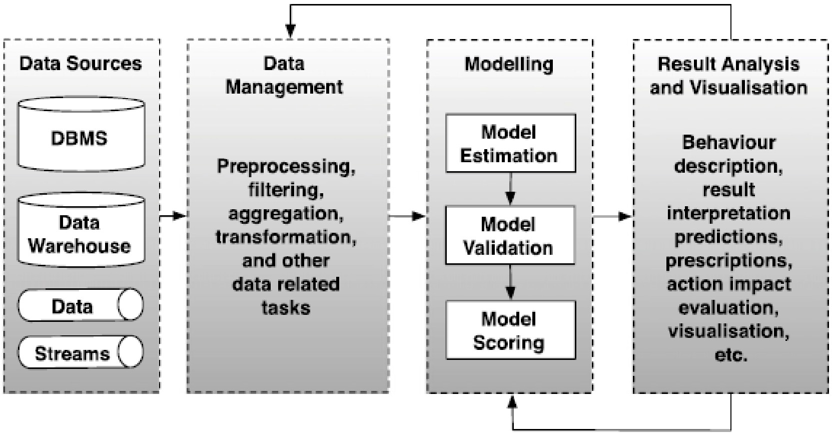

Data Sources
DBMS
Data Warehouse
Data
Streams
Data Management
Preprocessing, filtering, aggregation, transformation, and other data related tasks
Modelling
Model Estimation
Model Validation
Model Scoring
Result Analysis and Visualisation
Behaviour description, result interpretation predictions, prescriptions, action impact evaluation, visualisation, etc.

Fig. 8 Overview of the analitycs workfolw for Big Data

- ➢ Data variety and velocity
- ➢ Data storage
- ➢ Data integration solutions
- ➢ Data processing and resource management

2.5 Smart City

The concept of Smart City (SC) as a means to enhance the life quality of citizen has been gaining increasing importance in the agendas of policy makers. However, a shared definition of SC is not available and it is hard to identify common global trends.

Deakin defines the smart city as one that utilizes ICT to meet the demands of the market (the citizens of the city), and that community involvement in the process is necessary for a smart city. A smart city would thus be a city that not only possesses ICT technology in particular areas, but has also implemented this technology in a manner that impacts the local communty.

There are two main domains of Smart city which are considered according to the made investments. The first one is called Hard Domain and consists of the following sub-domains: energy Grid; public lightning, natural resources and water management; Waste Management; Environment; Transport, Mobility and Logistics; Office and residential buildings; Healthcare and Public Security.

Four sub-domains constructed the Soft main Domain of SC: Education and culture; Social inclusion and welfare; Public administration and (e-) government and Economy.

In most of the sources a smart system is associated with a digital platform. Thus the smart system must have suitable infrastructure, human capital and information. Consequently, in order to build smart economy one needs to have smart people, smart government, smart municipality, etc. The digital dimension has a strong relationship with intelligence and innovativeness (Komninos, 2006, 2011). The same source asserts that ICT is the main platform of knowledge-creating organizations and intelligent cities. 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, such as in Bletchley Park, where the Nazi Enigma cypher was decoded by a team led by Alan Turing. This has been referred to as the first example of a smart city or an intelligent community. [3]
- 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. There is much controversy surrounding this, particularly with regards to surveillance issues in smart cities.

In order to find the relation between a smart city and Big Data it is necessary to find a set of indicators strongly related with digital dimension. Within the next Section we are going to explore and define the linkage between Bid data and Smart Cities and how both conceptions affecting each other.

3 Smart City as generator of Big Data

The Linking element of Big Data and Smart Cities is the worldwide network of interconnected objects uniquely addressable based on standard communication protocols, or recently popular used concept Internet of Things.

A radical evolution of the current Internet into a Network of interconnected objects that not only harvests information from the environment (sensing) and interacts with the physical world (actuation/command/control), but also uses existing Internet standards to provide services for information transfer, analytics, applications, and communications. The Internet revolution led to the interconnection between people at an unprecedented scale and pace. The next revolution will be the interconnection between objects to create a smart environment. A schematic of the interconnection of objects is depicted in Fig. 9, where the application domains are chosen based on the scale of the impact of the data generated. [5]

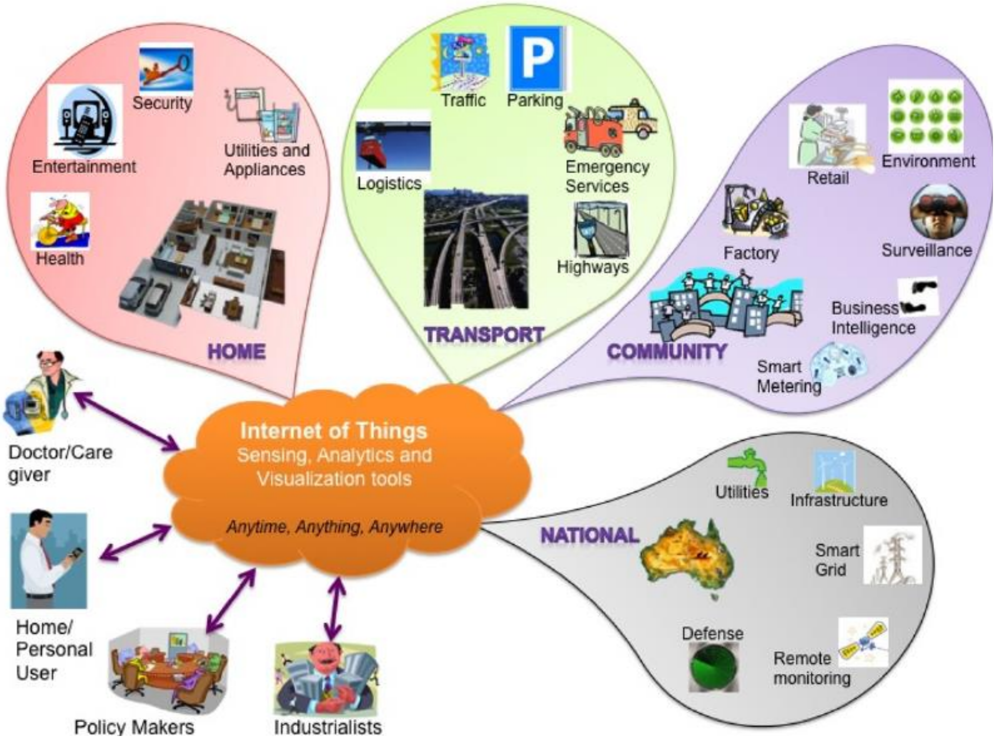


Fig. 9 IoT Schematic showing th end user and application areas based on data

Some devices transmit a few bytes of information at concrete times such as induction loop detectors for vehicle counting, while others generate heavy continuous data streams, such as CCTV surveillance. Therefore, smart cities are conceptually understood as systems of systems where the network infrastructure must handle different types of traffic. Consequently, to create a global overview of cities' data activity could become overwhelming due to its volume and heterogeneity [6].

Location intelligence is one of the relations between the data generating elements to their geographical location in cities. Generically, it is defined as the set of tools and methods that interrelates geographic information to business data, to identify patterns and relationships for decision-making that otherwise may be complex to operate with without a spatial representation [14]. As it can be seen, the digital dimension is distributed among the characteristics of smart cities. Thus it is possible to use the model of smartness that emphasizes this fact (Fig. 10).

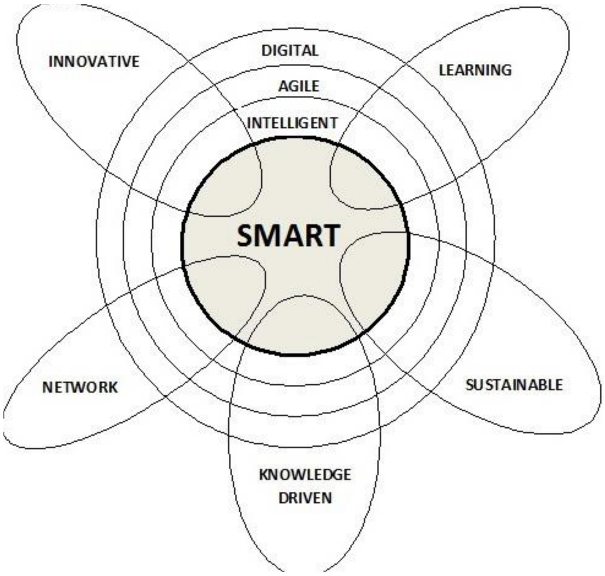


Figure 10.

4 Case study

In this section we are going to examine two different cases, in which data generating in the terms of Smart city ecosystem is determined by the mobility of the population and the geographical location of the studied cases.

Both cases concern the instrumentation intelligence of a smart city, as the first one is focused on Location intelligence and the second one examines the carbon emissions monitoring and imagery data integration. Bothe cases are in the range of smart city ecosystem.

Location intelligence may be useful to understand the interaction of data generating elements in relation to their geographical location in cities. Geographical Information Systems (GIS) combined with models abstracting the activity and mobility of data traffic sources as a function of well know city elements (households, roads....) may provide the appropriate spatio-temporal framework to approach the problem. In this way, dynamic data activity “heat maps” can be created to be able to visualize and understand the urban digital ecosystem. This happens with usage of Raw GIS data, which refers to basic GIS data as buildings and streets, as called in the case Basic Point (BP). Spatiotemporal data is represented by any set of data characterized by both a geographical location and time frame. Other element of the GIS is data activity, which refers to the digital data transmitted by devices within the smart city applications such as management of critical infrastructure, surveillance, etc. Three are the main types of entities which generate digital data, called for shortly Data point. So, they are: people, machines and vehicles .Very important for collecting and analysis of data is the Access to Network in the Smart City ecosystem. So in this case we are assuming that the city's access network is capable of handling all data transactions, in order to create the heat-maps. The basic idea of Geographical Information System is to model the behavior of all mentioned elements and to relate them to the available GIS data, instead of individual tracking and monitoring them. The idea proposed in this work is based on the quantification of the data volume that is being generated for any time period and at any point in the city, regardless of the entity and application used. The methodology is structured into three basic modeling levels, as illustrated in Fig. 11(a). [6]

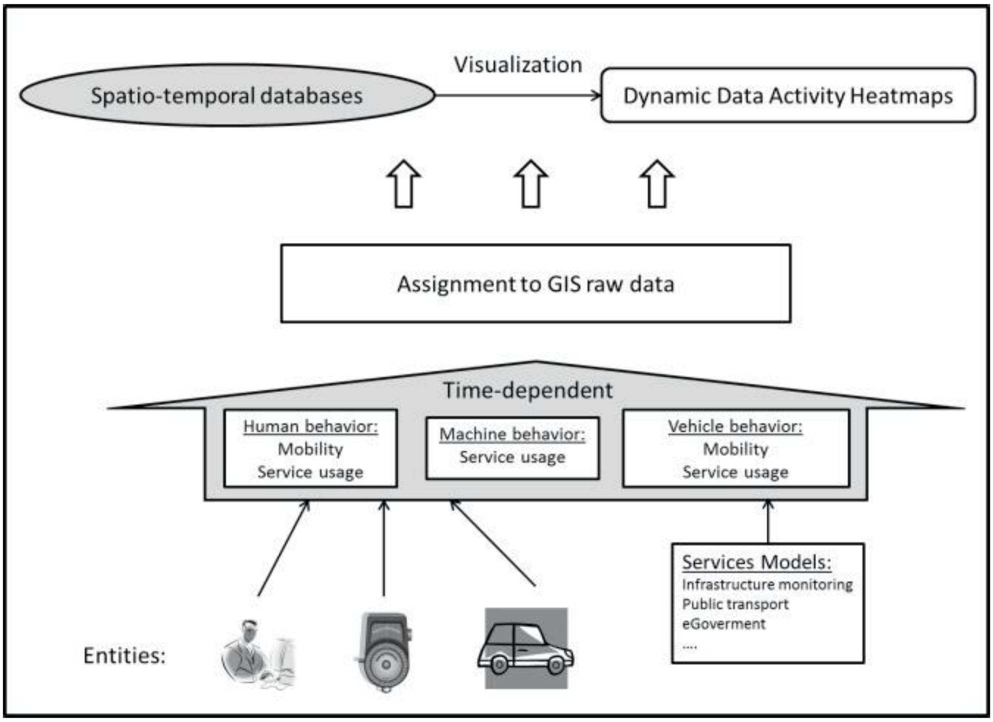


Fig 11a

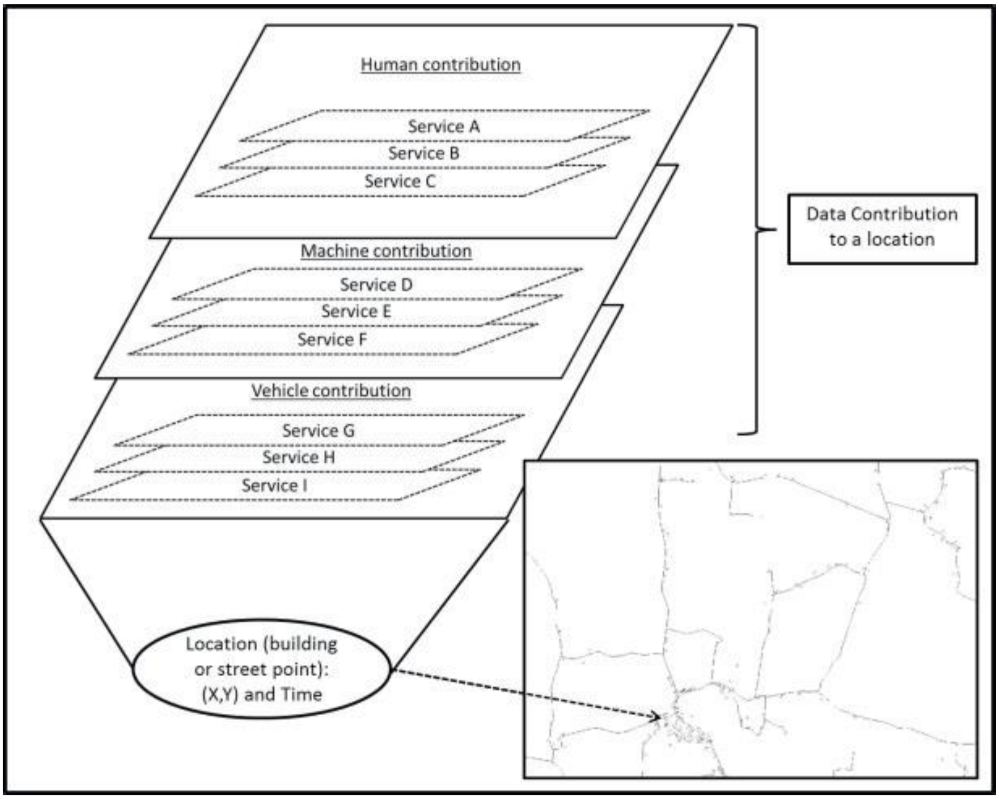


Fig. 11b

The methodology of GIS model is structured on three basic modeling levels: Entity, Service and Geo-coding levels. Entity level is related with the mobility of population. The Service modeling level concerns of the entities but the behavior patterns need to be extracted from the gathered data on Local and regional levels, thus the service levels depends on the specific location (factors and situation of the geographic area). Geo- coding Level derived from the combination of entity level and service level modeling to available GIS data (Basic Points, *BPs*). For example, in relation to Smart Grid, if a smart meter sends *X* kbytes every *Y* minutes and there is one meter per household, then it is possible to relate the data activity to a GIS *BP* (household in this case). Consequently, the models can be combined overlapping each other to estimate the data activity for each GIS element involved in the evaluation (in this work households and streets). Fig. 11(b) graphically illustrates the data contribution of different services and entities to *BPs*. In addition, to provide a uniform visualization layout, the area under study can be discretized into grid-cells where the data activity of each cell is the sum of the data volume of all the contained *BPs* at a given time period.

Second case which will be presented within this paper concerns a city in Korea. It is an ubiquitous city, newly developing as multi-functional administrative city. It has had already developed working systems – Carbon Emission Monitoring System (CEMS) and Closed Circuit Television (CCTV).

CEMS is designed and developed together with the urban information system (UIS) for municipal administration. On the basis of well-equipped urban ICTs, automatic remote meter-reading and data treatment system for carbon emission have been implemented in the U-Service concept. This system not only measures the quantity of emitted carbon but also identifies the characteristics of each household and house type by static and time-series statistics. To construct and use such kind of statistics on energy consumption is also very important to the decision-makers for making future policies. [7]

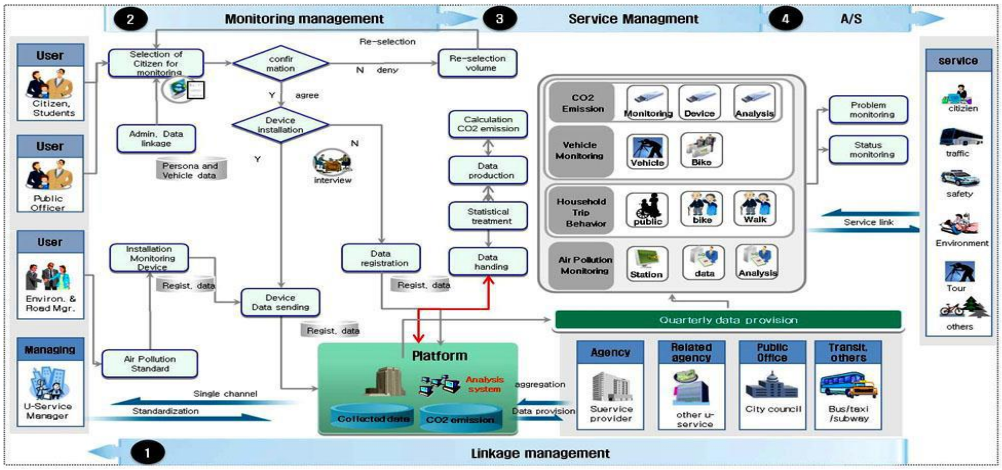


Fig. 12

CEMS provides 3 kinds of user-interfaces– basic data, cross-sectional data and time series data. The basic data interface presents and gets data by table format with which users can check their energy usage or input their own data by themselves. The cross-sectional data expresses by table and graphics. GIS techniques are adapted to generate the information map as a result of data processing. The time-series data is normally expressed in graph format. Daily, monthly data and trends over year can be presented via graph. The basic data and climate data are linked with UMS and other governmental agencies.

From examined two examples of GIS and CEMS can bring out common elements appearing in both, which basically generate the data nowadays - people (with their mobiles in hand and applications that have), their mobility and geo location. The Big Data generated by the first two factors, together with the processing of the third and appropriate and timely real time big data analysis, is sufficient and necessary condition for the design and development of each city to be turned into Ubiquitous and smart city. Minutely generated data by used ICTs and mobility applications contribute to the continuous development and innovative methods for processing, analysis and transformation of the enormous amount of big data, aimed of its beneficial use, to support the improvement and facilitation of life conditions in accretive density of urban areas.

5 Challenges and Trends

The volume of data operated upon by modern applications is growing at a tremendous rate, posing intriguing challenges for parallel and distributed computing platforms. These challenges range from building storage systems that can accommodate these large datasets to collecting data from vastly geographically distributed sources into storage systems to running a diverse set of computations on data. Some of the trends on Big Data Management and especially in the field of Analytics concern hardware platforms, software and application scope and visualization technologies. The Hardware platforms for data analytics trends are:

- Memory and Storage of Big Data
- Processing landscape for data analytics
- Network resources
- Energy Considerations in BD analytics

In addition to hardware enhancements, big-data analytics on yotta-scale datasets require a complete re-evaluation of the software stack. Improved software solutions should 1) scale to accommodate large datasets, 2) efficiently leverage the hardware platforms, and 3) bridge the increasing gap between the growth of data and computing power. The trends in the software are evolving in the storage system, data processing considerations and integration of different models. According to the scope of applications the trends of big data progressing will be in the following spheres:

- Health and human welfare
- Nature and natural processes
- Government and the public sector
- Commerce business and economic systems
- Social networking and the internet
- Computational and experimental processes [8]

Along with the trends there are some challenges to be opened for the big data. The mentioned ones below are not exhaustive, and as more research in this field is conducted, more challenging issues will arise.

Data variety: How to handle an always increasing volume of data? Especially when the data is unstructured, how to quickly extract meaningful content out of it? How to aggregate and correlate streaming data from multiple sources?

Data storage: How to efficiently recognize and store important information extracted from unstructured data? How to store large volumes of information in a way it can be timely retrieved? Are current file systems optimized for the volume and variety demanded by analytics applications? If not, what new capabilities are needed? How to store information in a way that it can be easily migrated/ ported between data centers/Cloud providers?

Data integration: New protocols and interfaces for integration of data that are able to manage data of different nature (structured, unstructured, semi-structured) and sources.

Data Processing and Resource Management: New programming models optimized for streaming and/or multidimensional data; new backend engines that manage optimized file systems; engines able to combine applications from multiple programming models (e.g. MapReduce, workflows, and bag-of-tasks) on a single solution/abstraction. How to optimize resource usage and energy consumption when executing the analytics application?[1]

6 Conclusions

Ubiquitous sensing enabled by Wireless Sensor Network (WSN) technologies cuts across many areas of modern day living. Cities around world are becoming “smarter” and regardless of the context of this evolution, digital data communications is one of the main common factors. In smart cities, a huge number of devices generate data traffic of very diverse nature associated to a large variety of applications. Also, the use of this enormous amount of data is very diverse ranging from infrastructure management to energy efficiency.

In this paper we reviewed some data models and techniques for data collection, extraction, processing and managing. Within the framework of Geographical Information System GIS is possible to combine and unify the data activity generated by different types of entities (people, vehicles…), using different applications and at different time periods. The method allows quantifying, geographically and dynamically in time, the volume of data generated to be used later for planning, managing, or monitoring purposes, among others. Moreover, the visualization of the resulting data activity may provide a convenient platform to understand the digital data dynamics in cities. Some of the biggest companies in the field of Big Data management and analytics are using common software to design their applications supported GIS. One of the basic elements in all of them is the programming model MapReduce. Such applications are Apache Hadoop (used also from IBM&Apple, VMWare) and HANA, developed by SAP.

References

[24] Marcos D. Assunзгo a,∗, Rodrigo N. Calheiros b, Silvia Bianchi c, Marco A.S. Netto c, Rajkumar Buyyab,∗ (2015) Big Data computing and clouds: Trends and future directions J. Parallel Distrib. Comput. 79–80 3–15

[25] Seema Maitreya*,C.K. Jhab (2015) MapReduce: Simplified Data Analysis of Big Procedia Computer Science 57 (2015) 563 – 571

[26] Komninos, Nicos. "Intelligent cities: Variable geometries of spatial intelligence.". In Deakin, Mark; Al Waer, Husam. From Intelligent to Smart Cities. Journal of Intelligent Buildings International: From Intelligent Cities to Smart Cities 3 (3).

[27] Amir Gandomi∗, Murtaza Haider. (2015) Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35 137–144

[28] Jayavardhana Gubbia, Rajkumar Buyyab,∗, Slaven Marusic a, Marimuthu Palaniswami a. (2013) Internet of Things (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems 29, 1645–1660

[29] Michael Jensen, Jose Gutierrez*, Jens Pedersen., Location Intelligence Application in Digital Data Activity Dimensioning in Smart Cities (2014)., Procedia Computer Science 36, 418 – 424

[30] Yountaik Leema*, Sang Ho Leea, Jungho Yoonb, Linking Data and Converging Systems for Smarter Urban Services: Two Cases of U-City Service in Korea, Procedia Environmental Sciences 22, 89 – 100

[31] Karthik Kambatlaa,∗, Giorgos Kollias b, Vipin Kumarc, Ananth Gramaa, Trends in big data analytics, (2014), J. Parallel Distrib. Comput. 74, 2561–2573

[32] Paolo Neirotti, Alberto De Marco, Anna Corinna Cagliano, Giulio Mangano ⇑, Francesco Scorrano, (2014)., Current trends in Smart City initiatives: Some stylised facts, Cities 38, 25–36

A Preferences Value Based Model Supporting Decision Making for a Smart Forest Ecological Management

Mariyana Lyubenova, Department of Ecology and Environmental Protection, Sofia University, Sofia, Bulgaria, ryana_l@yahoo.com

Yuri Pavlov, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria, yupavlov15@isdip.bas.bg

Alexandre Chikalanov, University of Library Study and Information Technology, Sofia, Bulgaria, ctmdevelopment@yahoo.com

Kamen Spassov, Faculty of Mathematics and Informatics, Sofia University, Sofia, Bulgaria, kspassov@gmail.com

Abstract: The multifunctional importance of forest ecosystems is often the reason for contradictions and conflicts between economic, social and environmental benefits. Diverse business interests in the forestry sector and the need to preserve the environment requires adequate decisions regarding the forest resources utilization and the sustainability of raw materials stocks. The sustainability of stocks should be provided by: forest resources monitoring; forest cultivation; consistent with resource preservation and regeneration felling; balanced and multifaceted use of forest as a source of both – material and other services, such as supporting, regulating and cultural. The main objective of complex investigation is to develop a value based model (VBM) for an integrated assessment of economic, ecological and social forest ecosystems services, which could be applied for monitoring and ecological management of forest areas and allowing the sustainable development of business and used forest resources. The modeling of such a complex system as the ecosystem is a difficult task, especially if you need its functions to be regulated and subordinated to the multifaceted sustainable use. For the modeling purposes we accept tree ecological indicators (factors), which describe adequately the main objective. These factors are: timber reserves in $m^3.ha^{-1}$ for the assessment of economic effects (material services); species richness, $n.ha^{-1}$, for the assessment of ecological effect (regulating and supporting services) and percentage of population employed in the forestry sector for the assessment of social effects or services. The model represents a multiattribute utility function, developed on the Decision making theory, Utility theory and stochastic approximation technic used as machine learning. The utility functions were calculated using the preferences of the Decision maker (DM), professional in forest ecology and environmental protection. The graphical representations of the functions are presented in the paper. The model will be used as a part of ForEco, a Smart Forest Ecological Management System that provides services to a large number of customers. ForEco is a Decision Support System (DSS), which will integrate a few models designated to support an adaptive ecological management of forest areas to achieve their sustainable development and operation as a natural source of ecosystem services. DSS will operate with a large quantity of Open Data. The software will be based on the FIWARE open software.

Keywords: Ecosystem Services, Forest, Mathematical Model, Decision Making, Smart Management.

1 Introduction

Ecosystem services and their economic evaluation

In general, the economic evaluation of the forest is limited to the revenue from timber, although the forest offers many different benefits. The concept of “multifaceted use of forests" has arisen since the mid-1970s [1]. Economic equations become more complicated because there is a need to recognize revenues from a number of simultaneous uses. Some of these alternatives contribute to the increase of revenues, if there is a balance between the benefits. For the first time Calish, Fight and Teeguarden (1978) included revenues from non-timber forest use in the equation of the rotation theory. They recorded five non-timber benefits: availability of game and fish, a variety of not hunting animals, visual aesthetics, erosion protection and water balance. For Douglas fir forest they assess non-timber benefits of 75% of total economic benefits [2]. For the first time the economic value of ecosystem services globally, incl. forests, was measured in the temperate climate zone (amounted to 894.10^9 dollars annually) by Costanza and a team of 12 authors [3], later it was evaluated by Convention on Biological Diversity (CBD) [4]. The works on the economic value of biodiversity was published by Pears [5], Harison et al. [6] and others. In 2001 the Secretariat of the CBD offered a methodology for evaluation of ecosystem services of forests, implemented and further developed in the Millennium Ecosystem Assessment (MEA) [7] and latest works [8]. In MEA, which served as a model for the next natural - economic evaluations, services of forest ecosystems were listed in the following order: food, wood, drinking water, fuel, regulation of water flow, regulation of diseases, carbon sequestration, regulation of local climate, medicine, recreation, aesthetics and spiritual values. Ecosystems as a whole provide the public services in the following four categories: material - energy, regulation, support, culture. Most summed the services that forest ecosystems provide are: recreation; climate change mitigation; biodiversity conservation; protection of soil, water and infrastructure; cultural aspects [9]. Forest products except timber, are grouped as edible plants and mushrooms, animal products, medicinal plants, hay, planting material and other (oils, resins, etc.). In 2009 Chinese scientists proposed to consider seven categories of forest services: conservation of water, soil retention, service protection / prevention, carbon sequestration and release of oxygen, biodiversity conservation, regulation of air quality, forest ecotourism. A study in the Beijing area showed that forest ecosystem services resulting from the existence of the forest, formed 87% of the annual economic value, the value of forest products was 11% and the share of sociocultural benefits was 2%. The authors of economics of ecosystem services study (TEEB project) proved that maximum use of one forest service negatively affected the state of the others and ultimately led to the economic decline in the final product. They have introduced the term “turning point”. “The impact factor accumulates without significant visible effect. After passing the turning point in the use of a service, the entire ecosystem degrades and other ecosystem services are lost” [10]. In 2012 a new platform IPBES has been established as the leading intergovernmental body for assessing the state of the planet's biodiversity, its ecosystems and the essential services they provide to society [11].

Multifaceted forest use

The concept of multiple benefits of forests was developed rapidly in the 1980s and received a strong impetus from the United Nations Conference on Environment and Development in Rio de Janeiro in 1992. The concept continues to evolve today. The focus on forestry has shifted from timber production to a wide range of goods and services (non-timber products, landscape design, conservation of biodiversity and more). New developments in this area are interested in an integrated assessment of forests through computer modeling. They have already measured not only the potential of forests, but also the risks associated with the large-scale multifaceted forest use [12 - 17].

Historically Bulgaria keeps pace with the global trend, developing the concept of forests with special uses. At the end of 1960 already applied the principle of continuous and uniform use within the increment, and then laid down the rules for the identification and protection of certain specific features of the forest. In the early 1990s the exaggerated use of wood is suspended [18, 19]. K. Bogdanov [20, 21] concludes that the multi-resource potential of Bulgarian forest can be divided into material and immaterial services. The author classified immaterial resources (services) of forests as environmental - protective, recreational and sanitary - protective. Because of its biodiversity, Bulgarian forests provide a rich opportunity to develop secondary uses, but the multifaceted management of forest resources is not yet an operational concept in Bulgarian forests. The majority of the proceeds of Bulgarian forests continue to come from the sale of wood. Lately the encouraging fact is that the scientific community in Bulgaria is working on mapping and valuation of ecosystem services, including forests, on Program BG03 "Biodiversity and Ecosystem Services" (Norwegian Financial Mechanism) [22, 23].

Contemporary accents of the problem

Today the concept of ecosystem services seeks to overcome the closed nature of the economic and environmental assessments, each of which is concentrated in his isolated range of problems and analytical apparatus and giving the general picture of material and energy flows defining the economy [2, 24]. The concept of sustainable development provides to profit from forests more than do at present. But profit is a consequence of multifaceted ecological management - the optimal solutions are only in the balance between the economic, social and environmental priorities of forest management. The using of forests, which is in conflict with this management, should not be tolerated. The modernizing of the forestry sector management should lie on private capital, which is not guided only by the desire for profit but on the concept of sustainable development. The other problem is that the ecological management of multifaceted forest use is difficult to implement in practice from "clean" foresters or private entrepreneurs only, without the presence of environmental experts or the relevant software products supporting the management through which will be realized sustainable development of forest resources and businesses.

Objective and tasks

The main objective of the paper is to present a complex investigation for developing of a value based model (VBM) for integrated assessment of economic, ecological and social services of forest ecosystems, which could be applied for monitoring and ecological management of forest areas and allowing the sustainable development of business and used natural resources.

The model is focused on the preferences of the expert in ecological management and the utility function. The model will be integrated within the ForEco system, described briefly. ForEco [25] is a Smart Forest Ecological Management System which will provide ecosystem services to a large number of customers based on open data and using FIWARE technologies [26]. The concept was born during the long term research activities of the team leader Dr. Lyubenova and team members, including the ones obtained as team members of FP6/FP7 projects and a number of actions under the COST Program. Team members were involved in the FP7 ELLIOT Project and developed an Internet of Things (IoT) Platform which is being adapted to a number of IoT applications. Later on this platform has been adapted to the FIWARE technologies and prepared for exploitation under the ForEco activities.

In addition to ecosystem models it was recognized that expert preferences can play crucial role in the decision making process. When experts are not present, decision makers can use appropriate Decision Making Support Systems in which experts' preferences are modeled given a particular objective.

2 Methods

2.1 Utility function, mathematical introduction

An analytical mathematical technique for the quantitatively description of complex economic, ecological and social sustainable use based on ecological management is applied.

The meaning of ***best*** varies from problem to problem [27]. The complexity of the wood biological and social processes, due to the inherent time variant properties and the lack of precise measurement make difficult the quantitative descriptions of the social human notions and ecological human expectations related with them. The determination of the optimal (the best) process parameters is also a challenge. The incomplete information could be compensated, with expression of qualitative human preferences in regard to the main purpose and the related sub-objectives as external manifestations of the human estimations. The necessity of a merger of empirical knowledge with mathematical exactness causes difficulties. Possible approach for solution of these problems is the stochastic programming and the Utility theory [27, 28]. The Utility theory basically deals with the expressed subjective preferences. Possible criteria for "the meaning of ***best***" can be an expert (decision maker-DM) utility function [28, 29, 30].

We need some mathematical formulations. Standard description of the utility function application is presented by figure (1).

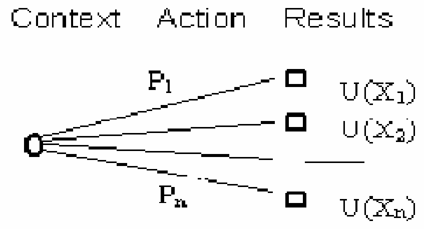


Fig. 1 Utility function application

There are a variety of final results that are a consequence of the expert or DM's choice and activity. This activity is motivated by a DM (Decision maker's) objective which possibly includes economic, social, ecological or other important process characteristics. A multiattribute utility function $U(x_i)$ assesses each of the final results (x_i, $i=1\div n$). The DM judgment is based on the DM choice and is measured quantitatively (discrete set of final results in this case) by the following formula [28, 29]:

$$u(p)=\sum_i p_i u(x_i),\ \ p=(p_1,p_2,..,p_i,..p_n),\sum_i p_i=1.$$

We denote with p_i subjective or objective probabilities which reflect the uncertainty of the final result. The strong mathematical formulation is the following. Let Z is a set of alternatives and P is a subset of discrete probability distributions over Z. A utility function is any function u(.) which fulfils:

$(p\succ q,\ (p,q)\in P^2)\Leftrightarrow(\int u(.)dp>\int u(.)dq),\ (p,q)\in P^2.$

The DM’s preference relation over P ($Z\subseteq P$) is expressed by($\succ$). Its induced indifference relation (~) is defined thus: $((x\sim y)\Leftrightarrow\neg((x\succ y)\vee(x\prec y)),\ (x,y)\in Z^2)$. We denote with ($\int u(.)dp$) integration based on the probability measure p. It is known that the existence of an utility function u(.) over Z determines the preference relation ($\succ$) as a negatively transitive and asymmetric one [26]. We mark the lottery “appearance of the alternative x with probability α and appearance of the alternative y with probability $(1-\alpha)$” as $<x,y,\alpha>$ [25, 27]. It is assumed that an utility function u(.) exists and that is fulfilled $((q,p)\in P^2\Rightarrow(\alpha q+(1-\alpha)p)\in P$, for $\forall\alpha\in[0,1])$. These conditions determine the utility function with precision up to an affine scale (interval scale), $u_1(.)\sim u_2(.)\Leftrightarrow u_1(.)=au_2(.)+b$, a>0 [25]. The following notations will be used:

$A_u=\{(\alpha,x,y,z)/(\alpha u(x)+(1-\alpha)u(y))>u(z)\}$,

$B_u=\{(\alpha,x,y,z)/(\alpha u(x)+(1-\alpha)u(y))<u(z)\}$.

The expected DM utility is constructed by pattern-recognition of A_u and B_u. This proposition is useful [30]:

PROPOSITION 1:

If $A_{u1}=A_{u2}$ then $u_1(.)=au_2(.)+b$, a>0.

The following description presents the procedure for evaluation of the utility functions. The DM compares the "lottery" $<x,y,\alpha>$ with the alternative z, $z\in Z$ ("better-$\succ$, $f(x,y,z,\alpha)=1$”, "worse-$\prec$, $f(x,y,z,\alpha)=-1$” or "can’t answer or equivalent-~, $f(x,y,z,\alpha)=0$”, f(.) denotes the qualitative DM answer). The DM relates the “learning point” (x,y,z,α)) to the set A_u with probability $D_1(x,y,z,\alpha)$ or to the set B_u with probability $D_2(x,y,z,\alpha)$. The probabilities $D_1(x,y,z,\alpha)$ and $D_2(x,y,z,\alpha)$ are mathematical expectation of f(.) over A_u and B_u, respectively, $D_1(x,y,z,\alpha)=M(f/x,y,z,\alpha)$, if $M(f/x,y,z,\alpha)>0$, $D_2(x,y,z,\alpha)=-M(f/x,y,z,\alpha)$, if $M(f/x,y,z,\alpha)<0$. Let $D'(x,y,z,\alpha)$ is the random value:

$D'(x,y,z,\alpha)=D_1(x,y,z,\alpha),\ M(f/x,y,z,\alpha)>0;$

$D'(x,y,z,\alpha)=-D_2(x,y,z,\alpha),\ M(f/x,y,z,\alpha)<0;$

$D'(x,y,z,\alpha)=0, M(f/x,y,z,\alpha)=0.$

We approximate D'(x,y,z,α) by a function of the type G(x,y,z,α)=(αg(x)+(1-α)g(y)-g(z)),where $g(x)=\sum_i c_i\Phi_i(x)$ and (Φ_i(x)) is a family of polynomials. Then the function g(x) is an approximation of the utility u(.).The function f(.) (DM answers) fulfills the conditions [30]:

f=D'+ξ, M(ξ/x,y,z,α)=0, M(ξ^2/x,y,z,α)<d, d∈R.

It is assumed that $u(x)\underset{L_2}{=}\sum_i r_i\Phi_i(x)$, r_i∈R, where (Φ_i(x)) is a family of polynomials. The following notations (see A_u) is used: t=(x,y,z,α), ψ_i(t)=ψ_i(x,y,z,α)=αΦ_i(x)+(1-α)Φ_i(y)-Φ_i(z). The stochastic algorithm described below realizes the evaluation procedure [30]:

$$c_i^{n+1}=c_i^{\,n}+\gamma_n\left[f(t^{n+1})-\overline{(c^n,\Psi\ (t^{n+1}))}\right]\Psi_i(t^{n+1})$$

$$\sum_n \gamma_n=+\infty, \sum_n \gamma_n^2<+\infty, \forall n, \gamma_n>0.$$

The coefficients c_i^n take part in the decomposition $g^n(x)=\sum_{i=1}^{N} c_i^n\Phi_i(x)$ and (c_i^n,ψ_i(t)) is the scalar product $(c^n,\Psi(t))=\alpha g^n(x)+(1-\alpha)g^n(y)-g^n(z)=G^n(x,y,z,\alpha)$. The line above $\bar{y}=\overline{(c^n,\Psi(t))}$ means $\bar{y}=1$ if y>1, $\bar{y}=-1$ if y<-1 and $\bar{y}=y$ if -1<y<1. The function G^n(x,y,z,α) is positive over A_u and negative over B_u depending on the degree of approximation of D'(x,y,z,α) and the function g^n(x) is the polynomial approximation of the empirical DM utility.

2.2 Evaluation of ecosystem services of preference – structuring of the main objective

The considered problem in a dialogue with the DM or expert is defined the following main objective of investigation: **to develop a value based model (VBM) for integrated assessment of economic, ecological and social forest ecosystems services, which could be applied for monitoring and ecological management of forest areas and allowing the sustainable development of business and use of forest resources**. Modeling of such a complex system as the forest ecosystem is a difficult task, especially if you need its functions to be regulated and subordinated to the multifaceted sustainable use. For the modeling purposes we accept tree indicators (sub-objectives or factors) adequately describing the main objective of investigation: X_1 - timber reserves ($m^3.ha^{-1}$) as representing criteria for the assessment of economic effects or material services; X_2 - species richness (n.ha^{-1}) as representing criteria for the assessment of ecological effect, or regulating and supporting services and X_3 - percentage of population employed in the forestry sector as representing criteria for the assessment of social effect or services. The model is developed as multiattribute utility function with the three factors mentioned. The coefficients of function were calculated using the preferences of ecology and environmental professional. Six graphical representations of the function with two fixed factors and one variable were calculated. The expert analysis and structuring carried out led to accepting the following sub-objectives and the appropriate criteria, which adequately describe the main objective and are real, physically measured quantities. We determine the domain of variation of representing criteria as follows:

- X_1-factor (material services as volume of timber per hectare - **economic effect**) [10 - 300 m^3.ha^{-1}];
- X_2- factor (regulating and supporting services - **ecological effect**) [1 - 200 number of species per hectare];
- X_3- factor (persantage of employed locals in the forestry sector - **social effect**) [1 - 30 %].

3 Rezults and Discussion

3.1 Utility evaluation and polynomial approximations

Graphically the structure of the main DM's objective has the form shown in figure 2. In the process of investigation independence by utility was found by the DM between the following factors:

- X_2 from X_1; X_2 from $X_{3;}$
- X_3 from X_1; X_3 from $X_{2;}$

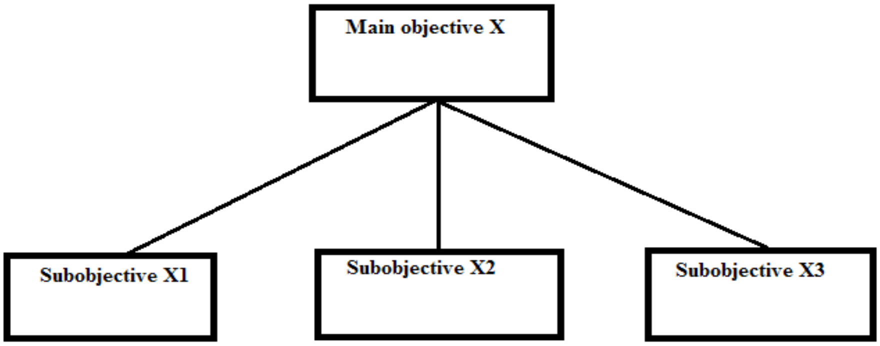


Fig. 2 Structure of the main objective and sub-objectives

The preferences of DM for x2 at different values of x1 and x3 do not change, suggesting independence of x2 from the changes of other two factors. Whatever the reserves of wood in the forest ecosystem are and the employment of the population in the forestry sector may be different, but in any case the preferences are directed to the presence of a large species richness of the forest, i.e. great variety of species of trees, grasses, moss, lichen, algae, animal species etc. that form the ecosystem and ensure its greater stability.

The preferences of DM for x3 at different values of x1 and x2 do not change, suggesting independence of x3 from the changes of other two factors. This means that whatever the reserves of wood are and at different species richness of the forest ecosystem, in any case, the preferences are aimed at the increasing of number of workers in the forestry sector. At low timber reserves and a poor species composition, the increasing of employment in forest sector is motivated by the need of reforestation, regeneration, cultivation of existing forests, optimization of forest-related natural resources, development of alternative uses and others. At high timber reserves and rich species composition, the increasing of employment is motivated by the opportunities of multifaceted use of forests and the need of environmental management and balanced utilization of forest resources.

Using the theory for decomposition of multiattribute utility to simpler functions given in (Keeney, 1999) we determine the following multiattribute utility structure:

$$u(X) = k_1 u(X_1; X^o{}_2; X^o{}_3) + \mathrm{f}_2(\mathrm{X}_1) \times \left[u(X^o{}_1; X_2; X^o{}_3)\right] + \mathrm{f}_3(\mathrm{X}_1) \times \left[u(X^o{}_1; X^o{}_2; X_3)\right] +$$
$$+ \mathrm{f}_{23}(\mathrm{X}_1) \times \left[u(X^o{}_1; X_2; X^o{}_3)\right] \times \left[u(X^o{}_1; X^o{}_2; X_3)\right], \text{ were } u(X^o{}_1; X^o{}_2; X^o{}_3) = 0$$
$$\text{and } u(X*_1; X*_2; X*_3) = 1.$$

In the formula above X^o=($X^o{}_1$; $X^o{}_2$; $X^o{}_3$)=(10,1,1) and X^*=($X^*{}_1$; $X^*{}_2$; $X^*{}_3$)= (300,200,30). The functions f_2, f_3 and f_{23} have the forms:

$$\mathrm{f}_2(X_1) = u(X_1; X*_2; X^o{}_3) - k_1 u(X_1; X^o{}_2; X^o{}_3),$$
$$\mathrm{f}3(X_1) = u(X_1; X^o{}_2; X*_3) - k_1 u(X_1; X^o{}_2; X^o{}_3),$$
$$\mathrm{f}_{23}(\mathrm{X}_1) = u(X_1; X*_2; X*_3) - \mathrm{f}_2(X_1) - u(X_1; X^o{}_2; X*_3).$$

Each of thise sixth functions was evaluated based on the DM’s preferences. For example the function u(X_1; $X^o{}_2$; $X^o{}_3$) has the form - figure 3:

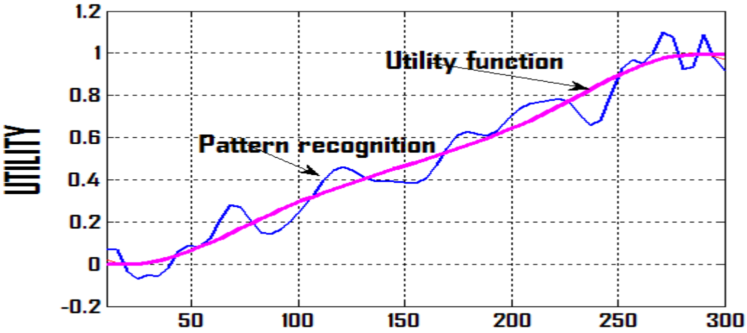


Fig. 3 Utility function $u(X_1; X^o_2; X^o_3)$

The blue seesaw line is pattern recognition of the positive or of the negative DM's preferences. The solid line is the evaluated Utility function polynomial approximation $u(X_1; X^o_2; X^o_3)$. In figure 5 is shown comparison betwen the evaluated Utility function $u(10; X_2; X_3)$ and the evaluated utility function $u(300; X_2; X_3)$.

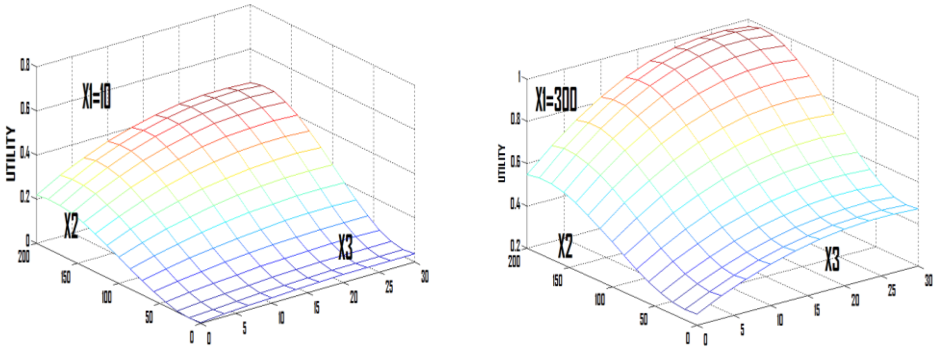


Fig. 4 Comparison between utility functions u(10;X$_2$:X$_3$) **and** u(300;X$_2$:X$_3$)

This utility model could be percieved as mathematical description of the social and ecological human expectations in the frame of the main objective of the complex wood problem.

3.2 Software solution

ForEco Smart Forest Ecological Management System will be built using FIWARE technologies. The proposed implementation will make use of the following open source components: *Generic enablers* (Orion Context Broker, EspeR4FastData, KeyRock Identity Manager and Application Mashup – WireCloud); *Open platform and Open source solutions* (CKAN Data Management, Opencart, JQuery – A Java Script library and MQTT Broker -

mosquitto MQTT real-time messaging broker). The general high level architecture of ForEco is presented on fig. 5.

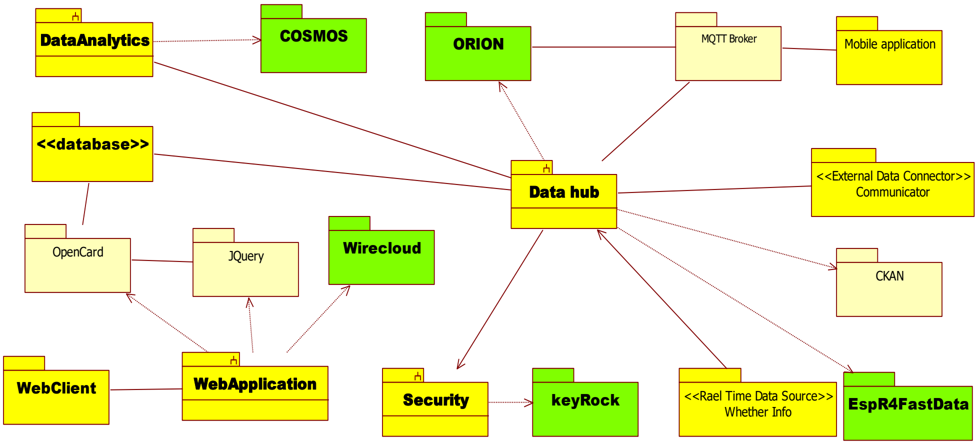


Fig. 5 ForEco general high lever architecture

4 Conclusion

A mathematical model was created to extract human knowledge and decision making patterns. The model was applied to simulate the decision making process of an ecology and environmental professional to maximize the utility function for a clearly defined goal. The described model will be included as a part of the Smart Forest Ecological Management System ForEco which is based on Open Data and FIWARE.

ForEco is a Decision Support System integrating a set of models that support adaptive ecological management of forest areas to achieve their sustainable development and operation as a natural source of ecosystem services. These ecosystem services are based on a large quantity of Open Data. The software will be based on FIWARE open software.

ForEco will give the opportunity to all interested parties (the society in general) to enjoy the benefits of healthy forests getting substantial economic effects. Ecological management of forests will lead to improvement of human health and at the same time will be an engine for new jobs. A scientific ecological management of the forest resources will provide a sustainable economic value in a long term not only in terms of revenues generated taking advantage of just selling the forest resources but also saving the costs to cure diseases, to provide fresh water, to clean the air, etc. Bio-waste utilization is another high added value business enabler. And last but not least a Smart Forest Ecological Management allows to a large number of society members to enjoy the recreational and cultural benefits.

Potential ForEco users are policy makers, governmental and municipal organizations (forest and landscape planners) which are involved in forest management activities, including planning new forests in an environmentally sound way with minimal negative effects on the environment; guiding forest managers and planners to select ecologically suited mostly native species to sites, instead of selecting a species and trying to modify the site to suit; economically effective resource management of forest timber; evaluate alternative strategies and providing comparable predictions; etc. Other prospective end-users are companies, clusters and experts from the Green Economy, Agricultural and Forestry sector, as well as – research and innovation organizations, innovative start-up and web-entrepreneurs, technology-transfer offices, and other stakeholders in the innovation chain. In addition, the ForEco platform will provide access to its services to citizens, educational organizations and NGOs who are interested and engaged in the project area of applications.

Acknowledgements: The research and the article are the results of joined efforts of a cross-functional team financed by the Finodex accelerator [31].

References

[1] Anderson, F.J. 1985. Natural Resources in Canada. Economic Theory and Policy. Toronto.

[2] Sabev, D. 2010. Humiliated land. Economic investigation. Sofia, "Demetri "OOD, 199 p.

[3] Costanza, R. a.o. 1997. The Value of the World's Ecosystem Services and Natural Capital. Nature, Vol. 387.

[4] The Value of Forest Ecosystems. Secretariat of the Convention on Biological Diversity. Technical series No. 4. 2001.

[5] Pearce, D., D. Moran. 1994. The Economic Value of Biodiversity. London.

[6] Harrison, P, P. Berry, G. Simpson, J. Haslett, M. Blicharska, M. Bucur, R. Dunforda, B. Egoh, M. Garcia-Llorente, N. Geamănă, W. Geertsema, E. Lommelen, L. Meiresonne, F. Turkelboom. 2014. Linkages between biodiversity attributes and ecosystem services: A systematic review. Ecosystem Services, 9, 191–203.

[7] Millennium Ecosystem Assessment. Ecosystems and Hyman Well – Being: Synthesis. Washington, 2005 - http://www.unep.org

[8] de Groot, R., L. Brander, S. Ploeg, R. Costanza, F. Bernard, L. Braat, M. Christie, N. Crossman, A. Ghermandi, L. Hein, S. Hussain, P. Kumar, A. McVittie, R. Portela, L. Rodriguez, P. Brinkm, P.van Beukering. 2012. Global estimates of the value of ecosystems and their services in monetary units. Ecosystem Services, 1, 50–61.

[9] European Forest Sector Outlook Study. UNECE / FAO, 2005, p. 94.

[10] TEEB – The Economics of Ecosystems and Biodiversity for Local and Regional Policy Makers. 2010, 18-23.

[11] Intergovernmental Platform on Biodiversity & Ecosystem Services – www.ipbes.net

[12] Nowak, David, D. Crane, J. Stevens, R. Hoehn, J. Walton, J. Bond. 2008. A Ground-Based Method of Assessing Urban Forest Structure and Ecosystem Services. Arboriculture & Urban Forestry. 34(6):347–358.

[13] Caldwell, P., G. Sun, S. McNulty, E. Cohen, J. Moore Myers. 2011. Modeling Impacts of Environmental Change on Ecosystem Services across the Conterminous United States.

[14] Sam Finn, Stephanie Keiffer, Becca Koroncai & Bob Koroncai. 2011. Assessment of InVEST 2.1 Beta: Ecosystem Service Valuation Software

[15] Volk, M. 2014. Ecosystem Services and River Basin Models: International Conference „Sustainability in the Water-Energy-Food Nexus“ C03 Special Session: Ecosystems and their Services in the Nexus, May 19-20, 2014, Bonn, Germany.

[16] Boumans, R., J. Roman, I. Altman, L. Kaufman. 2015. The Multi scale Integrated Model of Ecosystem Services (MIMES): Simulating the interactions of coupled human and naturalsystems. Ecosystem Services 12: 30–41.

[17] Lyubenova, M., A. Chikalanov, V. Lyubenova, St. Vatov. 2015. SPPAM 2.0 – Scientific description and Use. Journal of Environmental Science, Computer Science and Engineering & Technology, Sec. B, vol. 4, №1, 37-51.

[18] Sustainable Development of the Forest Sector in Bulgaria 2003 – 2013. Sofia, 2003.

[19] Stiptzov, V. 2006. Multifunctional forest management. Part 1. Multifunctional forest planning. Sofia.

[20] Bogdanov, K. 2002. Multifunctional forest management. Basis for regulation and sustainable development. Publishing University of Forestry, Sofia, p. 408.

[21] Bogdanov, K. 2008. Alarming Signs. Nature. Forests. Society. IK “Ruta” – HB, Sofia, p. 197.

[22] http://eeagrants.org/programme/view/BG03/PA02

[23] http://www.eeagrants.bg/en/2009-2014/

[24] Lyubenova, M., G. Gushlekov, A. Assenov. 2014. Need to Link the Two Crucial Strategies - Sustainable Development and Innovative Economy. Chemistry, Bulgarian Journal of Science and Education, 23, 3, 417-430.

[25] ForEco, www.foreco.net

[26] FIWARE, https://www.fiware.org/

[27] Keeney, R. & H. Raiffa. 1999. Decision with Multiple Objectives: Preferences and Value Trade-offs (2nd ed.). Cambridge & New York: Cambridge University Press.

[28] Fishburn, P. 1970. Utility Theory for Decision-Making. New York, Wiley.

[29] Raiffa, H. 1968. Decision Analysis. Introductory Lectures on Choices under Uncertainty. Addison-Wesley.

[30] Pavlov, Y., R. Andreev. 2013. Decision control, management, and support in adaptive and complex systems: Quantitative models, Hershey, PA: IGI Global.

[31] Finodex, www.finodex-project.eu

Bike Sharing System for Multimodal Mobility Based on Open Data

Roumen Nikolov[1], Galia Novakova[2], Elena Shoikova[1], Alexander Chikalanov[1]

[1] University of Library Studies and IT, Department of Computer Science
Boulevard "Tsarigradsko shose" 119, 1784 Sofia
r.nikolov@unibit.bg

[2] Sofia University, Department of Computing Systems, Faculty of Mathematics and Informatics
5 James Boutcher Str., 1164 Sofia, Bulgaria
g.novak@fmi.uni-sofi.bg

Abstract: The present paper aims at presenting methodology and intelligent techniques in the field of Big Data Management, Technologies and Applications within the scope of a currently ongoing project 'CloudBike'. The methodology of CloudBike setups an infrastructure, services and CloudBike Kits for cyclists in different user contexts and scenarios, such as: secure city cycling; cycling for wellbeing; digital social innovation campaigns for gathering user-generated data (UGD); etc. The CloudBike digital ecosystem consists of: 1) a cloud FIWARE (*fiware.org*) based infrastructure for collecting and processing data; 2) a cyclist's smartphone/tablet/PDA (computing and sens-ing device) with a mountable breakdown protecting and water-proof transparent case, and different sets of intelligent wearable sensors; 3) a mobile app that exploits both open data and UGD for provision of different (context based) services (including eCall). CloudBike makes possible implementation of comprehensive (bike-inclusive, user-centred) multimodal Smart City transportation systems and marketable products based on reuse of open data and UGD. It has potential substantial impact on other FINODEX domains.

Keywords: Mobile application, big data management, cloud technologies, multimodal mobility

1 Introduction

According to a report of the association of the European bicycle industry and the association of the *European two-wheeler parts' & accessories' industry* from July 2014, around 20 million bicycles are sold annually across Europe [1]. This total exceeds that of any other means of mobility (cars, motorcycles etc.). The statistics of the Association of European bicycle manufacturers Colibi [1] shows that the bicycle economics is on the rise with 3 of the largest markets being Germany (4.1 million bikes sold in 2014, +7.9%), France (2.98 million bikes sold in 2014, +7%) and the Netherlands (1.05 million bikes sold in 2014, +4.2%). Bulgaria is the sixth power in the production of bicycles among all countries in the European Union.

Another segment that has received wide popularity over the past few years is e-bikes (bikes with electric assistance), which increases with an average pace of around 30% on an annual basis [2]. The average price of bikes moderately increases and varies from EUR 844 in the Netherlands to EUR 307 in France.

This data reflects that currently there is a growing interest in cycling all over Europe, which is also a reflection of the growing interest to healthy life and protection of the environment. Currently, people are more and more willing to invest in (smart) bicycles that are constructed with new materials, which make it possible to cycle more, more comfortably and more safely at the same time.

In addition, the services based on the use of open data and UGD could be of use to a wide range of users – from city/municipal/national administrations mandated to optimize and plan the public transport, information providers (such as Google, Microsoft, Apple), public transport companies, etc.

Some existing smart bicycle applications are those of StravaLab (http://engineering.strava.com) [3]:

-*GPX to Road:* Upload a GPX file or a Strava activity and convert it to an editable route.

-*The Clusterer:* 100 million Strava activities clustered. Explores the worldwide top routes by distance and activity type.

-*Slide*: Auto-drawing map geometry from the Strava global dataset.

-*Roster:* Visually analyze athletic social habits, group activities, and preferred training partners.

-*Global Heatmap:* A total of 140 million rides and runs from Jan 2014 - May 2015, 375 billion data points visualizing the best roads and trails worldwide.

-*Routing Errors:* Route creation errors reported by Strava users are displayed for OSM map editors.

-*Top stops, Bay Area:* The best places to meet friends, have coffee, or just check out the view.

-*FlyBy:* Payback your activity along with those riding near you. Find who you passed or flew by while riding.

The user interface of CloudBike is very simplified: well readable text; maps; speech via a headset; speech generated information: navigation, messages, alerts; a pair of physical buttons to alternate a short list of (pre-programmed) modes of functionality. The Minimal Viable Product (MVP) will effectively exploit the Android based personal device functionalities with a small set of wellbeing smart sensors. Technical excellence will be ensured by a set of key performance indicators (KPIs), smart online community involvement; Living Lab, User Experience (UX) evaluation and gamification methods. The main product will be a mobile application named CloudBike able to interact with cloud FIWARE based infrastructure. CloudBike will be installed on a smartphone/tablet/PDA used by cyclists. It will enable them to get online access to different services in the context of multimodal smart city transportation systems.

The present paper is structured in five sections: After the Introduction, the general and specific objectives are highlighted in Section 2. Section 3 presents the open data in use for the realization of the discussed product. In Section 4 the main customers of the product are outlined. Finally, the paper draws some conclusions in Section 5.

2 Objectives

The general objective of CloudBike is to develop a FIWARE based infrastructure, open data and UGD based services and CloudBike Kits for cyclists in different user contexts and scenarios, such as: secure city cycling; cycling for wellbeing; child cycling; professional cycling; mountain biking; participation in digital social innovation campaigns; etc. The specific objectives are as follows:

a) to develop project management procedures, project platform and a CloudBike smart community;

b) to design, develop and deploy a CloudBike infrastructure and a set of basic open data and UGD based services;

c) to design, develop and deploy a CloudBike testbed and a Living Lab environment;

d) to design, develop and implement motivating experimental environment, scenarios, set of KPIs, gamification models, experiments and UX evaluation and data analytics methods;

e) to design and develop prototype of the MVP - CloudBike System and the CloudBike Kit for wellbeing;

f) to ensure and support a steady technological and business development and high quality of the MVP;

g) to develop a viable sustainability plan, a business model and a business plan, a dissemination and exploitation plan;

h) to do research on the product/service design, including by using Living Lab/ UX Design and UX Evaluation methodology;

i) to design a use-case and prototype.

3 Open Data in use

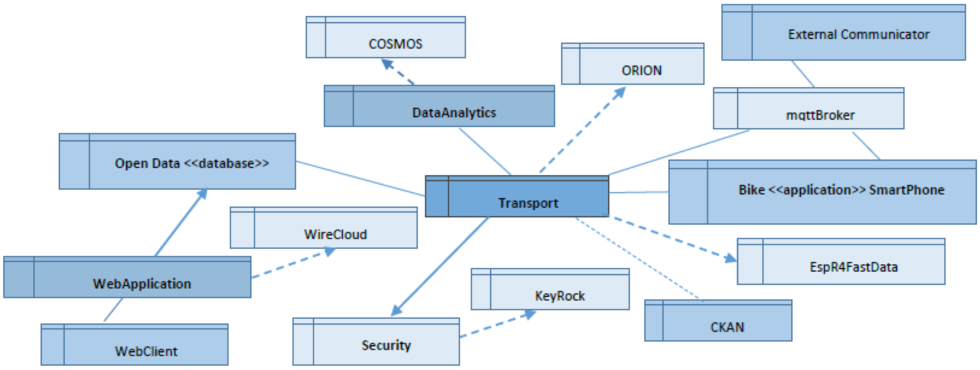


Figure 1. Basic architecture diagram of CloudBike with different FIWARE technologies

The role of the *Transport* subsystem will be to coordinate all data flows between participating parties. Sources of context awareness information can be any Open Data sources with the necessary license [4]. The Orion Context Broker will be used to implement context awareness for measurements coming from the field. In addition, the Esper4FastData enabler will be used to process data in real time. This subsystem will also integrate Open Data sources published by means of the *CKAN data management system*. Depending on different detected scenarios special events will be fired and an *Mqtt compliant broker* will notify all participants that are subscribed for them. A Bike Smartphone application will play a double role in the process as both a publisher and a subscriber. The package named *ExternalCommunicator* will allow external data publishers and subscribers to supply and get data by using a *REST protocol*. The incoming data will be propagated to all subscribed parties including the Data analytics subsystem and the database. The Security subsystem will be built on the *KeyRock Identity Manager*. The *Identity Manager* will be the central component that will provide a bridge between all subsystems at connectivity-level and application-level. All participants connected to *Transport subsystem* will have to pass an authorization & trust management by Single Sign On procedure carried on by *KeyRock*. The Data analytics subsystem will be based on extensive use of the *Cosmos* implementation of the *Big Data GE*. It will be used for the creation, use and deletion of private computing *Hadoop* clusters that will allow distributed processing of the accumulated datasets. To that end, a *Cygnus* type connector will be employed to transfer and store the context data from the *Orion Context Broke*r to *HDFS*. The end-user access to the system will be through a Bike Smartphone application and a Web application. The Web Application will be built on the base of the *WireCloud* implementation of the *Application Mashup GE*. *WireCloud* builds on cutting-edge end-user development based on Web application mashups. Web application mashups integrate heterogeneous data, application logic, and UI components (widgets/gadgets) sourced from the Web to create new coherent and value-adding composite applications. The proposed architecture will be implemented in an *OpenStack* environment - a free and open-source cloud computing software platform. *OpenStack* has a modular architecture with various code names for its components: *OpenStack Compute (Nova)* is a cloud computing fabric controller, which is the main part of an IaaS system. It is designed to manage and automate pools of computer resources and can work with widely available virtualization technologies as well as bare metal and High-Performance Computing (HPC) configurations.

4 Customers

The main customers of CloudBike will comprise individual bikers, respectively people who are already using bicycles or plan to start in the short run. Some part of this user group is using bikes for pleasure/sports and typical for the group is that people are active, take care of their health, they are interested in environmental issues and are keen in introducing new technologies at the same time. Another large user group of CloudBike consists of people using bicycles for transportation. Again, they are active, interested and open to new technologies and are mainly concentrated in cities where there is appropriate infrastructure for cycling.

Most of the future users of CloudBike will be people between 18 and 55 years who are both following the technological trends and can afford the price of the application. Among them the share of men is approximately as twice as high as the share of women. Another user group comprises members of biking clubs, organizations that provide bike rental (including municipalities), biking parks, etc.

The indicative price for the basic CloudBike application will vary from EUR 40 to 50 (silver and gold versions with more functionalities and options will also be developed). The application will have regular updates and will be offered as a downloadable product from the Internet. A general communication channel to reach the potential users of CloudBike will be social media (FB, Twitter, etc.) as a place to share social experiences since cycling could be referred as such. By using these media, the access to specific target groups all over the world will be made relatively easy. Creating a profile in LinkedIn, for example, will make possible to take advantage of the growing trend among the companies to be socially responsible or sensitive, which indirectly could form attitudes towards cycling and particularly intelligent cycling among its staff. By establishing the right contacts there the information about CloudBike will reach companies and people in key positions in them. At the same time, joining the thematic groups to discuss innovations and publication of relevant information will give additional credibility to the company.

Based on the estimated potential market of nearly 20 million bike users on an annual basis, it can be assumed that during the development phase at least 5,000 users will try the application as another group of 5,000 users is expected to test it during the tuning phase.

5 Conclusions

CloudBike will make it possible the implementation of comprehensive (bike-inclusive, user-centred) multimodal Smart City transportation systems and will additionally motivate people to use bicycles for transportation more intensively. By doing this, CloudBike will have positive effect on improving living conditions in cities - except the positive environmental effect from more intensive use of bicycles and the improved safety of citizens, there will be a significant economic effects too by reducing the time wasted in traffic jams and the lost profits. According to a report of the Joint Research Centre at the EC from 2012 on “Measuring Road Congestion”, road congestion costs an average 1% of EU’s Gross Domestic Product annually [5]. At the same time, CloudBike will stimulate the use of bicycles for health/sports/pleasure and will directly stimulate a more active lifestyle among the people in order to improve their general health status. As a cloud-based technology, CloudBike will be used for the development of marketable services based on reuse of open data and UGD. In future, indirect users of the information collected by the system can be, for example, city/municipal/national administrations mandated to optimize and plan the public transport; information providers (e.g. Google, Microsoft, Apple, etc.); public transport companies, etc. For example, the capital of Bulgaria (Sofia) has adopted a strategy for development of the bike transport which opens very good initial local market opportunities and support for the CloudBike product.

References

[1] Colibi. European Bicycle Market 2014 edition, Industry and Market Profile.

[2] European Cyclists’ Federation: www.ecf.com

[3] Smart bicycle applications of StravaLab: http://engineering.strava.com

[4] Open data portals - administrative, UGD and data generated from devices: http://s3platform.jrc.ec.europa.eu/open-data

[5] Christidis,P, J. Nicolás, I. Rivas (2012), Measuring road congestion, European Commission, Joint Research Centre, Institute for Prospective Technological Studies, doi:10.2791/15282, http://ipts.jrc.ec.europa.eu/publications/pub.cfm?id=5600

Automated Control Synthesis for Discrete Event Systems

František Čapkovič[*], Lyubka Doukovska[**], Vassia Atanassova[**]

[*]Institute of Informatics, Slovak Academy of Sciences,
Dúbravská cesta 9, 845 07 Bratislava, Slovakia
Frantisek.Capkovic@savba.sk

[**]Institute of Information and Communication Technologies, Bulgarian Academy of Sciences,
Acad. G. Bonchev St., Block 2, 1113 - Sofia, Bulgaria
l.doukovska@mail.bg, vassia.atanassova@gmail.com

Abstract: An approach to the automated control synthesis for discrete event systems (DES) is presented here. Petri nets (PN), namely the place/transition PN (P/T PN), are used here to model DES. First of all the approach suitable for the special kind of DES described by P/T PN named as state machines (SM) is presented. Then, the approach is extended for P/T PN with general structure. Namely, the reachability graph (RG) of such a system is utilized in order to find feasible state trajectories from a given initial state towards the prescribed terminal one. To avoid any confusion during the synthesis, renaming the transitions is necessary in order to obtain mutually different fictive names of the transitions. After completion the process of the synthesis, the fictive names of the transitions are renamed back to the original names.

Keywords: Control synthesis, directed graphs, discrete event systems, Petri nets.

1 Introduction

Behaviour of discrete event systems (DES) is discrete in nature. Namely, it depends on the occurrence of discrete events. In other words, DES are systems driven by discrete events. DES are widely used in human practice. Flexible manufacturing systems, different kinds of communication systems, transport systems of different kinds, etc. are typical representatives of DES. The behaviour of agents co-operating each other can also be understood to be a kind of DES. In [1], [2] causality of the DES behaviour was analysed. Petri nets (PN) of different kinds are often used to model DES. Namely, PN yield both the mathematical model and the graphical one. The PN-based model, more precisely the place/transition PN (P/T PN)-based model was also introduced in [1], [2], [3] as well as the method for modelling and control of DES. The method presented here is performed in virtue of the model of state machines (SM). It yields the space of feasible state trajectories from a given initial state to a desired terminal state as well as the sequences of transitions realizing the control. In order to solve the DES control synthesis problem, it is favourable to automate this process as soon as possible, even to achieve the fully automated control synthesis. Here, P/T PN will be understood to be bipartite directed graphs (BDG) defined e.g. [4], [6], although they are suitable only for the special kind of DES modelled by a class of P/T PN named the state machines (SM). SM are P/T PN where any PN transition has only single input place and only single output one. However, the main aim here is to extend the applicability of the approach to the wider class of P/T PN. Namely, in this paper the validity of the method will be extended for DES modelled by the P/T PN with general structure, where any PN transition can have (in contrast to SM) several input places and several output ones. Firstly, the real P/T PN with general structure is transformed into a fictive SM. Namely, the reachability graph (RG) of any P/T PN has the character of SM. Then, the control synthesis is performed by means of the fictive RG-based model corresponding to P/T PN model. Secondly, the fictive results are transformed back into the real notation. It is necessary to emphasize that the main aim here is to find existing paths from the given initial state of the P/T PN model of DES to the prescribed terminal one.

2 The P/T PN-based model of DES

The analytical model of P/T PN has the form of the following linear discrete system

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{B}.\mathbf{u}_k, \quad k = 0,1,\cdots,N \tag{1}$$

$$\mathbf{B} = \mathbf{G}^T - \mathbf{F} \tag{2}$$

$$\mathbf{F}.\mathbf{u}_k \le \mathbf{x}_k \tag{3}$$

where k is the discrete step of the dynamics development; $\mathbf{x}_k = (\sigma_{p_1}^k, \cdots, \sigma_{p_n}^k)^T$ is the n-dimensional state vector; $\sigma_{p_i}^k \in \{0,1,\cdots,c_{p_i}\}, i = 1,\cdots,n$ express the states of atomic activities of the PN places by 0 (passivity) or by $0 < \sigma_{p_i} \le c_{p_i}$ (activities); c_{p_i} is the capacity of p_i; $\mathbf{u}_k = (\gamma_{t_1}^k, \cdots, \gamma_{t_m}^k)^T$ is the m-dimensional control vector; $\gamma_{t_j}^k \in \{0,1\}, j = 1,\cdots,m$ represent occurring of the elementary discrete events (e.g. starting or ending the activities, failures, etc.) by 1 (presence of the corresponding discrete event) or by 0 (absence of the event); **B**, **F**, **G** are matrices of integers. More details can be found in [3] and/or in [1], [2].

3 The BDG-based model of DES

From the P/T PN based model we can derive the adjacency matrix of the bipartite directed graph

$$\mathbf{A}_{BDG} = \begin{pmatrix} \mathbf{0}_{n\times n} & \mathbf{F} \\ \mathbf{G} & \mathbf{0}_{m\times m} \end{pmatrix} \tag{4}$$

where the matrices **F**, **G** are the incidence matrices of the P/T PN model (1)-(3), however, being SM. The zero blocks with the designated dimensionalities express the fact that there are no edges either among the places each other or among the transitions each other. There exist directed edges only among the places and transitions (expressed by the incidence matrix **F**) as well as among transitions and places (expressed by the incidence matrix **G**). Then, the BDG-based model of SM can be expressed as follows. Setting $\mathbf{s}_k = \left(\mathbf{x}_k^T, \mathbf{u}_k^T\right)^T$ to be the (*n* + *m*)-dimensional state vector of BDG-based model of SM and setting and $\mathbf{\Delta} = \mathbf{A}_{BDG}^T$, we have

$$\{\mathbf{s}_{k+1}\} = \mathbf{\Delta}.\{\mathbf{s}_k\},\ \ k = 0, \cdots, 2N-1 \tag{5}$$

where $\{\mathbf{s}_{k+1}\}$ in general, is an aggregate all of the states that are reachable from the previous states $\{\mathbf{s}_k\}$. According to graph theory [4], [6], the maximal length of the path in a digraph *G* with *n* nodes is (*n* – 1) steps. It means that in our case for the length of a state trajectory is *N* ≤ (*n* - 1) steps. The straight-lined development of the SM dynamics (expressed by BDG-based model) from the initial state $\mathbf{s}_0$ towards a terminal state $\mathbf{s}_k$ is

$$\{\mathbf{s}_k\} = \begin{cases} (\{\mathbf{x}_{k/2}\}^T, \mathbf{0}_m^T)^T, \text{ if } k = 0, 2, 4, \ldots, 2N-2 \\ (0_n^T, \{\mathbf{u}_{(k-1)/2}\}^T)^T, \text{ if } k = 1, 3, 5, \ldots, 2N-1 \end{cases} \tag{6}$$

where $\mathbf{0}_j$ in general, is the *j*-dimensional zero vector;

$$\mathbf{x}_{k/2} = \mathbf{G}^T.\mathbf{u}_{(k-2)/2},\ k = 2, 4, \ldots, 2N-2,\quad \mathbf{u}_{(k-1)/2} = \mathbf{F}.\mathbf{x}_{(k-1)/2},\ k = 1, 3, \ldots, 2N-1. \tag{7}$$

Store the particular vectors as the columns of the matrix

$$\mathbf{M}_1 = \left(\mathbf{s}_0,\ {}^1\{\mathbf{s}_1\}, \ldots,\ {}^1\{\mathbf{s}_{2N-1}\},\ {}^1\{\mathbf{s}_{2N}\},\right) \tag{8}$$

The principle of the method consists in the intersection of both the straight-lined development of the BDG-based model (5) from the given initial state $\mathbf{x}_0$ towards the desired terminal state $\mathbf{x}_t$ and the backward development of the following model, where the system matrix is the transpose of the system matrix $\mathbf{\Delta}$ in (4)

$$\{\mathbf{s}_{2N-k-1}\} = \mathbf{\Delta}^T.\{\mathbf{s}_{2N-k}\},\ \ k = 0, \ldots, 2N-1 \tag{9}$$

and the structure of the vectors is the same. Then,

$$\mathbf{M}_2 = \left({}^2\{\mathbf{s}_0\},\ {}^2\{\mathbf{s}_1\}, \ldots,\ {}^2\{\mathbf{s}_{2N-1}\}, \mathbf{s}_{2N} \right) \tag{10}$$

After the *intersection* of the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ when we consider that

$\{\mathbf{s}_i\} = \min({}^1\{\mathbf{s}_i\}, {}^2\{\mathbf{s}_i\}),\ i = 0, \ldots, 2N$ and ${}^1\{\mathbf{s}_0\} = \mathbf{s}_0$, ${}^2\{\mathbf{s}_{2N}\} = \mathbf{s}_{2N}$ we obtain

$$\mathbf{M} = \left(\mathbf{s}_0, \{\mathbf{s}_1\}, \ldots, \{\mathbf{s}_{2N-1}\}, \mathbf{s}_{2N} \right) \tag{11}$$

Here, information about the state vectors is stored in the even columns of **M** while information about the control vectors is stored in its odd columns.

4 Extending the approach for P/T PN with general structure

The above mentioned BDG-based approach to the control synthesis handles the PN incidence matrices **F**, **G**. However it is possible solely in case of P/T PN being SM. Namely, SM keeps the PN transitions of special kind – with the single input place and the single output place. In P/T PN with general structure several places can occur at the input and/or at the output of any transition. Fortunately, the RG of P/T PN can be understood to be SM and the algorithm given in Fig. 1 can be applied.

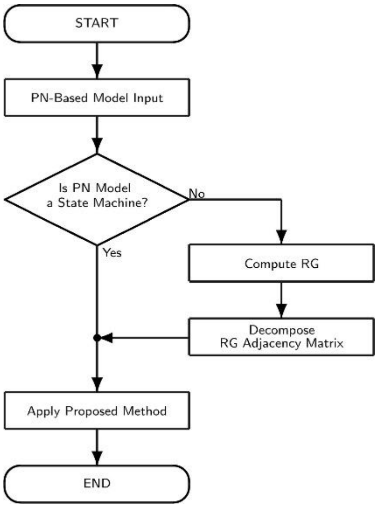


Figure 1. The algorithm for the generalization

However, the RG corresponding to the P/T PN with the general strycture does not contain the matrices **F**, **G**. Namely, in such a case RG is characterized by both the adjacency matrix **A** (which has not the structure (4)) or better said by the quasi-functional adjacency matrix $\mathbf{A}_k$ with entries given by indices of transitions fixed to RG edges between the RG nodes and the matrix $\mathbf{X}_{reach}$ storing in its columns the P/T PN feasible states. Therefore, let us disassemble the matrix $\mathbf{A}_k$ into the matrices $\mathbf{F}_{RG}$, $\mathbf{G}_{RG}$. However, there exists still one obstacle or difficulty on this way. Namely, the original P/T PN transitions among the entries of $\mathbf{A}_k$ usually occur more frequently than once. Consequently, confusion could occur during the computational process and decline it. To avoid these difficulties, it is necessary to rename the original P/T PN transitions before disassembling $\mathbf{A}_k$. In such a way we obtain the mutually different fictive transitions. Each of them occurs in the fictive adjacency matrix $\mathbf{A}_{Tr}$ only once. Namely, the number of the fictive transitions is Tr being the global number of the $\mathbf{A}_k$ entries different from zero. The renaming of the $\mathbf{A}_k$ entries is performed raw-by-raw so that the non-zero elements are replaced by integers - the ordinal numbers starting from 1 and finishing at Tr. Thus, the auxiliary matrix $\mathbf{A}_{Tr}$ is obtained. Its structure and dimensionality is the same like the structure of $\mathbf{A}_k$. Only its nonzero entries are different from those in $\mathbf{A}_k$. The disassembling of the matrix $\mathbf{A}_{Tr}$ into the incidence matrices $\mathbf{F}_{RG}$, $\mathbf{G}_{RG}$ is accomplished as follows. Their elements depends on the transformation matrix $\mathbf{T}_{Tr}$ between the real and fictive transitions as follows

$$\mathbf{T}_{Tr}(\mathbf{A}_k(i,j), \mathbf{A}_{Tr}(i,j)) = \begin{cases} 1, \text{ if } \ \mathbf{A}_k(i,j) \neq \mathbf{0} \ \& \ \mathbf{A}_{Tr}(i,j) \neq \mathbf{0} \\ \qquad 0, \text{ otherwise} \end{cases} \tag{12}$$

$$\mathbf{F}_{RG}(i, \mathbf{A}_{Tr}(i,j)) = \begin{cases} 1, \text{ if } \ \mathbf{A}_k(i,j) \neq \mathbf{0} \ \& \ \mathbf{A}_{Tr}(i,j) \neq \mathbf{0} \\ \qquad 0, \text{ otherwise} \end{cases} \tag{13}$$

$$\mathbf{G}_{RG}(\mathbf{A}_{Tr}(i,j), j) = \begin{cases} 1, \text{ if } \ \mathbf{A}_k(i,j) \neq \mathbf{0} \ \& \ \mathbf{A}_{Tr}(i,j) \neq \mathbf{0} \\ \qquad 0, \text{ otherwise} \end{cases} \tag{14}$$

$$i = 1, \ldots, n_{RG}; \ j = 1, \ldots, n_{RG}$$

We can say that now we have the fictive SM with the incidence matrices $\mathbf{F}_{RG}$, $\mathbf{G}_{RG}$ which has the same RG like the original P/T PN with general structure. Setting the transpose of the fictive adjacency matrix of BDG to be

$$\mathbf{\Delta} = \mathbf{A}^T_{BDG} = \begin{pmatrix} \mathbf{0}_{n_{RG} \times n_{RG}} & \mathbf{G}^T_{RG} \\ \mathbf{F}^T_{RG} & \mathbf{0}_{T_r \times T_r} \end{pmatrix} \tag{15}$$

we can utilize the control synthesis procedure proposed for SM. After finishing the control synthesis, the names of the transitions have to be transformed again, in order to retrieve their original names. Using the fictive state vectors instead of the real ones as well as the fictive control vectors instead of the real ones we obtain (from the matrix $\mathbf{M}$)

$$\mathbf{X} = (\mathbf{x}_0, \{\mathbf{x}_1\}, \ldots, \{\mathbf{x}_{N-1}\}, \mathbf{x}_N); \quad \mathbf{U}^* = (\mathbf{u}_0, \{\mathbf{u}_1\}, \ldots, \{\mathbf{u}_{N-1}\}); \, \mathbf{U} = \mathbf{T}_{T_r} . \mathbf{U}^* \qquad (16)$$

where the symbol * means that the renamed (fictive) transitions were used in the computational process. The results related to the original transitions are computed by the transformation matrix.

4.1 The illustrative example

Let us illustrate here how the P/T PN model with the general structure (which cannot be characterized as SM in any case) can be transformed into the SM. Consider P/T PN-based model displayed on the left in Fig. 2. It is clear that the P/T PN is not any SM. Moreover, it contains double directed arcs, namely from p_1 to t_1, from p_4 to t_2 and from t_3 to p_4. Its structural parameters are the following

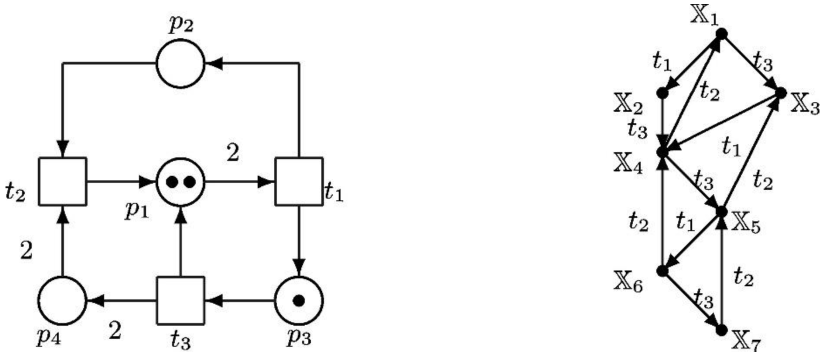


Figure 2. The P/T PN-based model of DES (left) and the corresponding RG (right)

$$\mathbf{F} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 2 & 0 \end{pmatrix}; \mathbf{G} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 \end{pmatrix}; \mathbf{B} = \begin{pmatrix} -2 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & -2 & 1 \end{pmatrix}$$

For the given initial state $\mathbf{x}_0 = \begin{pmatrix} 2 & 0 & 1 & 0 \end{pmatrix}^T$ we obtain the RG of the P/T PN model. The RG parameters are expressed by the following matrices

$$\mathbf{A}_k = \begin{pmatrix} 0 & 1 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \end{pmatrix}; \mathbf{X}_{reach} = \begin{pmatrix} 2 & 0 & 3 & 1 & 2 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 2 & 2 \\ 1 & 2 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 2 & 2 & 4 & 4 & 6 \end{pmatrix}$$

The non-zero entries of $\mathbf{A}_k$ represent the indices of the transitions assigned to the corresponding edges from the node N_i to the N_j of the RG represented by the element $\mathbf{A}_k$(i,j). The columns of

$\mathbf{X}_{reach}$ represent in Fig. 2 the feasible state vectors $\mathbf{X}_j$, j=1, ..., 7, being, as a matter of fact, the nodes $\mathrm{N_j}$ of the RG. To transform the general P/T PN into SM (with respect to Fig. 1) let us rename the transitions. It means that we have to renumber the nonzero entries of the matrix $\mathbf{A}_k$ to remove the multiplicity of the entries. The renaming of transitions assigned to the RG edges between corresponding RG nodes (i.e. the indices being the nonzero entries of $\mathbf{A}_k$) is made row-by-row. Because the number of nonzero entries of $\mathbf{A}_k$ is $T_r = 11$, the transformed adjacency matrix and the transformation matrix are as follows

$$\mathbf{A}_{T_r} = \begin{pmatrix} 0 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 7 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 9 & 0 & 0 & 10 \\ 0 & 0 & 0 & 0 & 11 & 0 & 0 \end{pmatrix}; \mathbf{T}_{T_r} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Consequently, the fictive incidence matrices of the fictive SM are the following

$$\mathbf{F}_{RG} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{G}_{RG}^{T} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

In such a way we obtained the fictive SM model corresponding to the original P/T PN model. Now the procedure (4) - (11) suitable for synthesizing the SM control can be used. Thus, we can obtain the final results, namely, the state trajectories as well as the control trajectories. While the control trajectories are the causes, the corresponding state trajectories are their consequences.

$$\mathbf{X}=\begin{pmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{pmatrix};\ \mathbf{U}^*=\begin{pmatrix}1&0&0&0&0\\1&0&0&0&0\\0&1&0&0&0\\0&1&0&0&0\\0&0&0&0&0\\0&0&1&0&0\\0&0&0&0&0\\0&0&0&1&0\\0&0&0&0&0\\0&0&0&0&1\\0&0&0&0&0\end{pmatrix};\ \mathbf{U}=\begin{pmatrix}1&1&0&1&0\\0&0&0&0&0\\1&1&1&0&1\end{pmatrix}$$

From these matrices it follows that there are two state trajectories

$$\mathrm{X}_1\to\mathrm{X}_2\to\mathrm{X}_4\to\mathrm{X}_5\to\mathrm{X}_6\to\mathrm{X}_7;\quad \mathrm{X}_1\to\mathrm{X}_3\to\mathrm{X}_4\to\mathrm{X}_5\to\mathrm{X}_6\to\mathrm{X}_7$$

Simultaneously, we have the following results of the control synthesis yielding the control trajectories which, as a matter of fact, are the cause of the state trajectories. For the fictive transitions the results are the following

$$t_1^*\to t_3^*\to t_6^*\to t_8^*\to t_{10}^*;\quad t_2^*\to t_4^*\to t_6^*\to t_8^*\to t_{10}^*$$

For the real transitions the trajectories are given by the reverse renaming the transitions, i.e.

$$t_1\to t_3\to t_3\to t_1\to t_3;\quad t_3\to t_1\to t_3\to t_1\to t_3$$

The final results with the real state trajectories and the real control ones can also be expressed graphically - see Fig. 3.

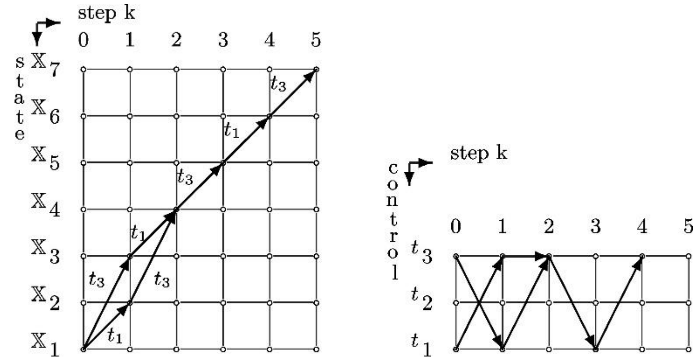


Figure 3. The final results. The state trajectories (left) and the control trajectories (right).

From the previous illustrative example it is clear that the P/T PN model with general structure can be transformed into the fictive SM model. Then, the control synthesis algorithm suitable for SM can be applied. Afterwards, the results achieved by means of SM model can be transferred back into the original denotation corresponding to the original P/T PN model.

5 Conclusions

The topic of the paper is the automated DES control synthesis modelled by P/T PN. As it is known form galaxy of papers, RG corresponds to any P/T PN. There is no ambition here to propose a new method for finding RG. On the contrary, the known RG (corresponding to the P/T PN-based model) is utilized here for finding the state trajectories (sequences of the P/T PN state vectors) from a given initial state to a prescribed terminal one. The BDG-based approach to DES control synthesis based on P/T PN was proposed and presented in this paper. While the method is directly suitable only for the special kind (as to the structure) of P/T PN named SM, having the transitions with single input place and single output one, after the transformation process the proposed method is suitable also for P/T PN with general structure where the transitions can have more input places and/or more output ones. The proposed method utilizes the RG of the general P/T PN as its SM model and the decomposition of its adjacency matrix into the fictive matrices, being the incidence matrices of the fictive SM. After such a procedure the BDG-based approach to the control synthesis is applicable. The approach was illustrated on the example.

The approach can be applicable not only for DES described by P/T PN with general structure, but also for P/T PN models of DES with modular structure as well as for cooperation of agents in MAS (multi agent systems).

Acknowledgement

The research work was performed in the framework of the joint project “Modern Paradigms in the Field of Intelligent Systems (2015-2017)” between Slovak Academy of Sciences and Bulgarian Academy of Sciences as well as in the framework of the VEGA project under grant # 2/0039/13.

References

[1] Čapkovič F. (2007) Modelling, Analysing and Control of Interactions Among Agents in MAS. *Computing and Informatics*, 26, 5, Slovak Academy of Sciences, 507-541.

[2] Čapkovič F. (2014) Cooperation of Agents in Complex Systems Based on Supervision. *Cybernetics and Information Technologies*, 14, 1, Bulgarian Academy of Sciences, 40-51.

[3] Čapkovič F., Doukovska L., Atanassova V. (2014) Comparison of Two Kinds of Cooperation of Substantial Agents, In: R.D. Andreev, ed., *Proceedings of the International Conference on Big Data, Knowledge and Control Systems Engineering*, Sofia, Bulgaria, 97-106.

[4] Diestel R. (2005) *Graph Theory*, 3rd Edition, Springer-Verlag, Heidelberg-New York.

[5] Murata T. (1989) Petri Nets: Properties, Analysis and Applications, *Proceedings of the IEEE*, 77, 4, 541-588.

[6] Bang-Jensen J., Gutin G. (2008) *Digraphs: Theory, Algorithms and Applications*, 2nd Edition, Springer-Verlag, London.

Knowledge and Data Management in "Civil Protection for ALL" project

Nina Dobrinkova, Valentina Terzieva, Katia Todorova

Institute of Information and Communication Technologies – Bulgarian Academy of Sciences

Acad. Georgi Bonchev bl. 2, 1113 Sofia, Bulgaria

e-mail: nido@math.bas.bg, valia@isdip.bas.bg, katia@isdip.bas.bg

Abstract: In most cases of emergencies or series of events that could lead to disasters, the needs of specific social groups (in particular children, families and elderly people) are not taken into account and served appropriately. This happens either because there are no special procedures and plans for them, and therefore there exist no knowledge and practical skills in the intervention teams, or because different priorities has prevailed so far, such as the general optimization of the urgent situation or local improvements that leave no room for specific interventions.

The media and society put a lot of pressure to professionals and volunteers involved in emergency management operations and even partial failure is not acceptable. Therefore, when assessing the intervention priorities, actions related to vulnerable social categories must be duly justified and examined several times before they are executed.

In our paper we will present how the CP4ALL project define the existing knowledge gaps when proffessionals have to deal with vulnerable groups on the field and we will present the best practices and tools envisaged for preparation actions in civil protection activities optimizing the care on the field for the vulnerable social groups in case of emergencies.

Keywords: CP4ALL project, vulnerable groups, emergency reactions.

1 Introduction

The project CP4ALL (Civil Protection for all) has been developed starting from real needs expressed by the Civil Protection organizations and local authorities part of the consortia core group that have touched with their hands the luck of specialized operators officially trained to deal with kids and other vulnerable groups on the field in cases of disaster situations. The idea of new approaches and trainings for the emergency teams has let to development of the project Civil Protection for ALL oriented to the volunteer groups helping and supporting kids and vulnerable groups on the field in case of major disaster affected their life.

The project main activities are grouped as follows:

1. Identify best practices and lessons learnt related to emergency preparedness towards the involvement of vulnerable categories in Europe and beyond, starting from previously financed prevention and preparedness projects.
2. Training sessions design, elaboration and implementation dedicated to the creation of specific teams of operators, specialized in vulnerable categories of citizens (small groups coming from different parts of one territory that will train other people after their training).
3. Establishment of a dynamic list of project and topic specialized professionals and volunteers, as well as design and development of a dynamic web based tool (application – online platform) to be used in case of emergencies in relation to vulnerable social groups' needs and challenges to be tackled. The platform will remain available to each partner's territory and beyond through disseminating activities and networking with civil protection and vulnerable groups related entities networks in Europe.
4. Planning, organization and evaluation of small scale pilot exercise that will implement project results in a realistic emergency simulation scenario and evaluate them.
5. Awareness-raising and dissemination campaign for people employed in sectors such as health care of vulnerable people, staff of civil protection, volunteers' organization etc.

All five envisaged activities should lead to increased capacity of organizations working with children and families on the field after major disasters. Thus will naturally result in optimizations when it comes to emergency situations and emergency preparedness through concrete actions that will enable organizations to plan, test, evaluate, adapt, adopt, and propose improvements for legislative and procedural improvements by state authorities.

2 Best practices identification

In the first phase of best practices identification the CP4ALL consortia used a questionnaire initially based on similar projects in the past from Interreg Programme, as that kind of projects deal a lot with best practices and the way they can be indispensable part of entities' policies and procedures. Psychologists that deal with kids gave also an initial feedback, as well as partners of CP4ALL. The collected information gave data on the source of good practice for every partner that provided it, its place and time of implementation, achieved goals and objectives, involved stakeholders and on which phase of its implementation the proposed best practices. [1,2,3,4,5,6,7,8]

The CP4ALL project partners collected 25 best practices:

Slovenia: 4
Bulgaria: 2
Greece: 4
Spain: 11
Italy: 4

Best practices have been implemented within each country; however few of them have been implemented in other countries. Most refer to activities implemented in regular basis and can be adapted in case of emergency. However few deal with past events management lessons learnt and transfer best practices from everyday working pace to emergency situation challenges. The context of each country best practices has been summarized in the analysis by covering the main bullet points listed:

- ✓ Role of Children in case of an emergency (not only as the subject of caring actions but also their ability in problem solving among them and sharing of feelings)
- ✓ Role of public entities (civil protection and supporting ones)
- ✓ Role of stakeholders from public sector, NGOs and private sector (e.g. camp owners)
- ✓ Role of parents and grandparents
- ✓ Role of Media (including social ones) and what is their role (how to provide information, to protect children, to give practical info/ guidance/ to raise awareness etc)
- ✓ Role of schools and other structures that provide education/ training/ moral values/ self estimation
- ✓ Role of Volunteer organizations, scouts etc.

The achieved result from this best practices collection was list of sources (entities and documents) that will have public use, and can be updated by all members of civil protection community, vulnerable social groups related organizations, as well as from research and academic institutes. Such a tool will help bridge the two worlds of civil protection and vulnerable social groups related organizations, by providing easy to use and reliable information on challenges, progress, issues and other elements that influence successful cooperation.

3 Educational platform used in CP4ALL for best practices dissemination and learning tool

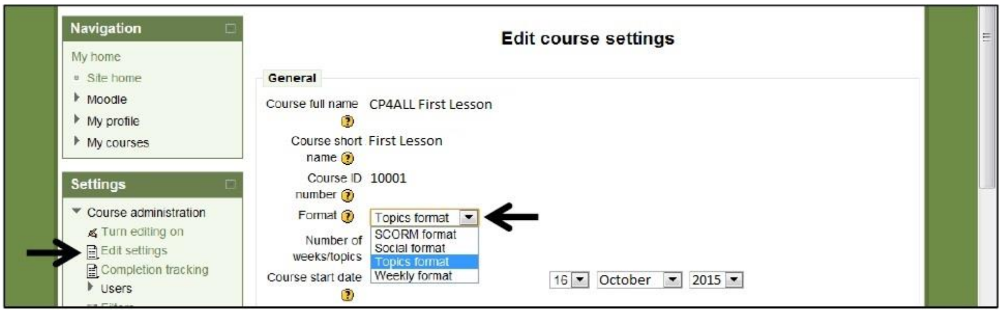
Globalization and the rapid ICT spread urge traditional classroom model to gives way to new forms of training - online, blended, mobile or distance supported by the use of computers or mobile devices. Thus, the tutors need to change their teaching approaches, to improve their digital literacy and to work with new technologies. In Civil Protection for All (CP4ALL) project one of the main tasks is to provide distance courses to trainees in different countries by implementing elements of best practices pool created within the project. In this regard an appropriate online training platform that supports the pedagogical provision of e-learning and the achievement of set objectives have to be selected. The key parameters that need to be taken into account are the open web based architecture and numerous people to use it. Besides, the platform has to allow expansion, differentiation and further development. In this context we need a comprehensive fully operational system that services teaching along with the possibility of providing opportunities for course publication, e-Assessment and e-Questionnaire fulfilment, various educational contents: specific textbooks, electronic exercises on a certain course, multimedia tools for a given topic, even educational online games. Interactive communication tools that support intuitive feeling for classroom learning and information exchange will benefit the users. In addition the multinational nature of the project demands for local languages usage concerning special features and contributions (learning materials, images, audios, videos, etc.)

Various existing platforms are evaluated and the best practises and past projects results are considered. Moodle – Modular Object Orientated Distance Learning Environment fully satisfies the project requirements. It is a free open source, user-friendly, available for many operational systems, flexible and feature-rich learning management system (LMS). It is a web-based platform that provides education for users distributed across the map as CP4All-users are. Because of the modular structure, it supports a lot of functions and customization. Taking into account the anytime/ anywhere accessibility Moodle offers an effective training delivery method and ensures self-education process (lack of direct contact between teachers and students). The main features are as follows:

- **Lesson –** enable many ways of creating, reusing and managing lessons and courses. The authoring tool allows learning resources with different formats (doc, pdf, ppt, etc.), to add sound, image, video, flash, etc. and to include events and activities.

- **Quiz and Test –** enable many approaches to evaluate learners' knowledge and to question them. The system supports creating several types of tests and questions (multiple-choice, true-false, short answer, computational, questions benchmarking, etc.). The questions can be kept in a database and reused in the same or different course if appropriate. The test can be evaluated automatically.
- **Wiki –** enable collaboration by several functions for group-working and help learning process.
- **Glossary –** provide instant information concerning terms in the course subject area.
- **Forum –** enable discussions among all registered users about educational issues.
- **Chat –** enable talks about any topics in real time with classmates and instructors.
- **Calendar –** enable scheduling lessons, tests, future commitments, events, etc.

Moodle is functionally–rich and based on the social constructivism pedagogy, so it can support standardized curriculum and interactive training. Besides that the platform allows a social community to be built, where users can discuss, interact, and work collaboratively. The courses can be organised on SCORM, topic, social or weekly basis (Fig. 1). The weekly format arranges the classes in week schedule with lessons, assignments, tests, etc. The topic format allows thematic or titled grouping of resources. The social format concerns the forum and is appropriate for discussions. This course format organisation contribute to the consistency and helps to prevent learners from knowledge gaps.



Edit course settings

General

Course full name: CP4ALL First Lesson

Course short name: First Lesson

Course ID number: 10001

Format: Topics format

Number of weeks/topics

Course start date: 16 October 2015

Figure 1.

The lecturers can create online courses, tests, assessments, publish any texts and illustrative materials, etc. and can fully control their settings, including permitting or restricting the access; they can define start/ end date and password for subscription. During the teaching process tutors can upload/ download files (lessons, manuals, texts, learners' work), give assignments, coursework or projects for individual or group study, run examination materials, use assessment tools, etc. Teachers can define their own grading scales and have an opportunity to decide whether and how to display learners' scores for assignments, tests and quizzes (word/ excel spreadsheet file). In addition they have some administrative functions such as students' registering/ deregistering and grouping; forum opening, messaging, seeking feedback through surveys (answers can be anonymous), etc.

Moodle provides full user logging and tracking so that tutors can monitor learners' activities (reading lesson, making assignment, writing in the forum, doing test); last access; number of times read lesson/ done test ; etc. as well as a detailed "snapshot" of each student involvement and time spend in the system. This activity reports can be presented with details in graphs and for each student or about each resource/ course. Thus, the gathered log-term information can be processed further with learning analytical Big Data tools. Learning analytics give tutors learning trends in order to enhance education and general reporting allows stakeholders and managers to see overall progress parameters. These reports are an essential component of the e-learning process both for teachers and learners. Moodle has standard reporting functions for educators (Grader, User, Overview, Activity Completion, and Course Completion Progress reports) and site reports (Activity Log, Change Log/ Config Changes, Course Automated Backups reports).

4 Implementation

Main objective of the CP4ALL project is to design and elaborate training sessions, concerning civil protection procedures, dealing with vulnerable social groups in emergencies and best practices that have been identified and roved in other cases. Key element will be the concept and operation of families and kids friendly spaces, suitable also in cases for sending/ receiving assistance to/from neighboring countries, in cooperation with national competent authorities. This concept will expand to other social groups such as elderly, handicapped, minorities. Such approach improves also the European joint mechanism intervention in emergencies. Training sessions will be joint events, consisting of personnel dealing with vulnerable groups, civil protection professionals, volunteers, scientists etc.

For evaluation of the volunteer groups selected by the partner's selection national exercises will take place on every territory and for evaluation of the results will be used an international training camp in Italy. At this small scale table top exercise every volunteer will follow the materials organized in the Moodle classrooms and will be evaluated and tested by civil protection professionals monitoring their progress during the course.

The planned lecturing hours during the table top exercise are 40 and as a result is expected that the challenges in civil protection field can deeper the cooperation between the volunteers and the civil protection authorities.

Civil Protection is a field that due to common concerns (protection) by all involved stakeholders, CP4ALL proposed approach can give successful results. CP4ALL project actions will contribute towards that direction. However, to be realistic, more things need to be done. Activation of the network of vulnerable social groups working entities that are being involved at civil protection activities in one or another way. Encouragement of civil protection authorities (at National, Regional and Local levels) to include special plans (or elements in already existing plans) and procedures about those social groups and how they will cooperate sufficiently in all phases of civil protection cycle. Establish integration of such activities in territories where people know each other and have a better sense of community and can work more closely together. Raise awareness of European civil protection community on all parameters that affect vulnerable groups during emergencies so that in future guidelines and directives on how to be prepared on such issues. Furthermore, research institutes could transform and transfer knowledge from other fields related to the problems addressed by CP4ALL project into civil protection domain knowledge and provide the necessary research evidence that will allow cooperation between civil protection authorities and vulnerable social groups related entities to produce best results. Such efforts can be further enhanced through specialized exercises of all kinds that will address those issues in the best way, at no expense of the overall effectiveness of civil protection forces. In that framework, also, volunteer organizations that already help civil protection professionals can have a significant role as they entail in their structure and functioning characteristics of both worlds. All these further steps must be represented not only in future plans but also in procedures followed in every step, to optimize preparedness and subsequently, management, restoration and prevention. Other initiatives can be used: Indicatively, LIFE+ for raising awareness and dissemination campaign, Interreg initiatives for bilateral or multilateral cooperation, ERASMUS for learning environments etc.

5 Conclusions

Bridging the worlds of civil protection and of vulnerable social groups (especially kids, families, elderly) caring entities is an issue that has not been tackled in a systematic way in the past. Only isolated approaches, partially discuss the problem and uncompleted efforts not only in Europe but worldwide have been performed. Therefore, the main objective of identifying the sources of best practises and lessons learnt on the issue, to map them, create a pool of best practices and select the most suitable ones for implementation on partners’ territories through web platform and training activities on small scale exercise is of great value. The use of best practices can be useful in civil protection field in support of vulnerable social groups. Although the project actions will end as scheduled, their results can be used further even after CP4ALL project termination.

6 Acknowledgement

This work is supported by the CP4ALL project co-financed by DG ECHO/SUB/2014 /695754. For more information web site: http://www.cp4allproject.eu/.

References

[1] http://www.aap.org/en-us/advocacy-and-policy/aap-health-initiatives/Children-and-Disasters/Pages/default.aspx?nfstatus=401&nftoken=00000000-0000-0000-0000-000000000000&nfstatusdescription=ERROR%3a+No+local+token.

[2] http://www.aap.org/en-us/advocacy-and-policy/aap-health-initiatives/Children-and-Disasters/Pages/Strategic-Plan-for-Disaster-Preparedness.aspx

[3] http://cybercemetery.unt.edu/archive/nccd/20110427002908/http://www.childrenanddisasters.acf.hhs.gov/index.html

[4] http://cybercemetery.unt.edu/archive/nccd/20110427005326/http://www.childrenanddisasters.acf.hhs.gov/reports_studies/resources/Preparedness_Indicators_Modified_from_CPG_101_2_v5.pdf

[5] http://cybercemetery.unt.edu/archive/nccd/20110427002913/http://mdchhs.com/sites/default/files/pdf_articles/JEM-9-2-01-children-and-disaster-planning.pdf

[6] https://www.childwelfare.gov/management/disaster_preparedness/

[7] http://muskie.usm.maine.edu/helpkids/rcpdfs/copingwithdisasters.pdf

[8] http://www.savethechildren.org/site/c.8rKLIXMGIpI4E/b.6206913/k.68FA/Preparing_for_Emergencies.htm

ISSN: 2367 - 6450

Institute of Information and Communication Technologies - Bulgarian Academy of Sciences

"John Atanasoff" Union on Automatics and Informatics, Bulgaria