Information Systems & Grid Technologies

Tenth International Conference ISGT'2016

Sofia, Bulgaria, Sep. 30 – Oct. 1., 2016.

ST. KLIMENT OHRIDSKI UNIVERSITY PRESS

ISGT’2016 Conference Committees

Chair

Prof Vladimir DIMITROV

Program Committee

- Mícheál Mac an AIRCHINNIGH, Trinity College, University of Dublin
- Pavel AZALOV, Pennsylvania State University
- Irena BOJANOVA, University of Maryland University College
- Marco DE MARCO, Catholic University of Milan
- Milena DOBREVA, University of Malta
- Vladimir GETOV, University of Westminster
- Seifedine KADRY, American University of the Middle East, Kuwait
- Kalinka KALOYANOVA, University of Sofia “St Cl Ochridsky”
- Angelika KOKKINAKI, University of Nicosia
- Violeta MANEVSKA, University of Bitola “St Cl Ochridsky”
- Maria NISHEVA, University of Sofia “St Cl Ochridsky”
- Dov TE’ENI, Tel-Aviv University
- Stanislaw WRYCZA, University of Gdansk
- Fani ZLATAROVA, Elizabethtown College

Organizing Committee

- Vasil GEORGIEV
- Maria KOLEVA

Vladimir Dimitrov, Vasil Georgiev (Editors)

Information Systems & Grid Technologies

Tenth International Conference ISGT'2016

Sofia, Bulgaria, Sep. 30 – Oct. 1., 2016.

Proceedings

organized by

Faculty on Mathematics and Informatics.
University of Sofia St. Kliment Ohridski

Bulgarian Chapter of the
Association for Information Systems (BulAIS)

St. Kliment Ohridski University Press

Preface

This conference was being held for the ninth time in the end of September, 2016 in the Rector's meeting hall of the University of Sofia "St. Kliment Ohridski, Bulgaria. It is supported by the Science Fund of the University of Sofia "St. Kliment Ohridski" and by the Bulgarian Chapter of the Association for Information Systems (BulAIS).

Total number of papers submitted for participation in ISGT'2016 was 25. They undergo the due selection by at least two members of the Program Committee. This book comprises 21 papers of 17 Bulgarian and 9 foreign authors. The conference papers are available also on the ISGT web page http://isgt.fmi.uni-sofia.bg/ (under «Former ISGTs» tab).

Responsibility for the accuracy of all statements in each peer-reviewed paper rests solely with the author(s). Permission is granted to photocopy or refer to any part of this book for personal or academic use providing credit is given to the conference and to the authors.

The editors

ISSN 1314-4855
St. Kliment Ohridski University Press

TABLE OF CONTENTS

Semantic Search in Heterogeneous Digital Repositories:
Technological Aspects
Maria Nisheva-Pavlova 7

Formal Specification of CAPECs in UML & OCL
Vladimir Dimitrov 15

Roboethics: questions related to the implementation of robots
Ioannis Patias 22

Modelling of GIS Based Visual System for Local Educational Institutions' Stakeholders
Snezana Savoska, Valmir Muaremi, Andrijana Bocevska,
Blagoj Ristevski and Zoran Kotevski 28

Constructing Phylogenetic Networks and Trees for the Analysis of Genetic Datasets
Elitsa Kaloyanova 40

Gamification Systems to Support the Learning Process
Teo Petrov, Kalinka Kaloyanova 48

COBBIS system cataloging – review
Blagojce Najdovski, Violeta Manevska, Snezana Savoska 58

Detailed Formal Specification of Software Weakness CWE-128
(Analyzes)
Vladimir Dimitrov 62

Detailed Formal Specification of Software Weakness CWE-128
(Specification)
Vladimir Dimitrov 70

Integrating software applications for analysis of big heterogeneous data
from parallel sequencing in a distributed cloud environment
Yana Nikolova, Vladimir Dimitrov, Luca Piredu, Dimitar Vassilev 78

The necessity of Project Management capacities and Organizational
Flexibility for the successful implementation of complex projects in Municipalities
Ioannis Patias 89

Management of heterogeneous data collections from external sources through PostgreSQL FDW
Svetoslav Savov, Vladimir Dimitrov, Dimitar Vassilev 97

Separation of Concerns in Motion Path Control of Redundant Robot Arms
Evgeniy Krastev 106

Formal Specification of CAPECs in CSP
Vladimir Dimitrov 120

Bulgarian e-Government Information System Based on the Common Platform for Automated Programming – Requirements
Ivan Stanev, Maria Koleva 125

Bulgarian e-Government Information System Based on the Common Platform for Automated Programming – Technical Solution
Ivan Stanev, Maria Koleva 134

Hierarchical clustering in evolutionary analysis – data preparation and validation
Dilyana Hadzhimateva, Irena Avdjieva, Valeriya Simeonova,
Dimitar Vassilev 152

Information Extraction from Articles Related to Crime Events
Lyubka Genova 158

A SAT Approach for Task Assignment
Vesela Angelova 170

System SlopeStabBG for Calculation of Stability Factor in Opencast Mines
Mariana Trifonova 178

Software Engineering Management Education – Bringing Industry Standards to the University Program
George Sharkov and Maya Stoeva 184

AUTHOR INDEX 200

Semantic Search in Heterogeneous Digital Repositories: Technological Aspects

Maria Nisheva-Pavlova

Faculty of Mathematics and Informatics, Sofia University St Kliment Ohridski
5 James Bourchier blvd., Sofia 1164, Bulgaria

marian@fmi.uni-sofia.bg

Abstract. Semantic search extends the scope of traditional search and information retrieval paradigms. The paper discusses the goals and the main characteristics of semantic search and analyzes the implementation of most popular semantic search systems. The presentation is focused on a number of experiments with free software tools for advanced semantic technologies accomplished within the Master's degree course in Semantic Technologies at the Faculty of Mathematics and Informatics, Sofia University.

Keywords: Semantic web, metadata, ontology, semantic annotation, semantic search, information retrieval, semantic interoperability

1 Introduction

Semantic search is aimed at improving the effectiveness and accuracy of traditional search relying on understanding the user's purposes and the right meaning of the particular terms and phrases that appear in the searchable repository or Web space. Semantic search systems use various information resources and techniques from different domains to provide adequate search results: dictionaries of synonyms, thesauri, ontologies, context of search, intent, location, generalized and specialized queries, natural language queries, etc. The implementation of a considerable part of existing semantic search engines is based on augmentation of the user queries with the aid of proper ontologies and dictionaries of synonyms while others are oriented to building and utilization of various semantic annotations using also available subject ontologies.

During the last decade, modern semantic technologies enable the rapid development of new software for semantic search and information retrieval. A significant part of these technologies are supported by corresponding free software or at least by restricted free versions of powerful commercial software tools.

The paper discusses the main characteristics of semantic search and the implementation principles of most popular semantic search systems. The implementation solutions realized within a project directed to building tools for semantic search in a particular heterogeneous digital library are presented in

V. Dimitrov, V. Georgiev (Editors): ISGT'2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

brief. The last section analyzes the results of a series of experiments with free software tools for advanced semantic technologies accomplished in the form of course projects within the Master's degree course in Semantic Technologies at the Faculty of Mathematics and Informatics, Sofia University.

2 Main Characteristics of Semantic Search

Semantic search extends the scope of traditional search and improves its results using an understanding of the meaning of the search term(s) given by the user. Semantic search augments the goals of traditional information retrieval, which is mostly oriented to document retrieval, with additional entity and knowledge retrieval tasks [Guha, 2003]. It improves the effectiveness of conventional search and information retrieval methods by taking into consideration the meaning of words that can be formalized and then described using ontology languages like RDF(S) and OWL [Wei, 2008]. In this way, a semantic search system is able to retrieve adequate results by reasoning on the user query and the accessible explicit domain knowledge.

Five main directions of research and development in semantic search have been outlined in [Mäkelä, 2005]: augmenting traditional keyword search with semantic techniques, basic concept location, complex constraint queries, problem solving, connecting path discovery.

Augmenting traditional keyword-based search with semantic techniques is based on the use of proper domain ontologies or thesauri as sources for the user query expansion. In particular, the popular WordNet (https://wordnet.princeton.edu/ wordnet/) and YAGO (http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/) ontologies are often used for this purpose. First, the concepts referred by the given keywords are located in the ontology; then, a partial traversal of the graph representing the ontology structure is accomplished in order to find the terms, semantically related to the discovered concepts. These terms are utilized to either broaden or constrain the search.

Most semantic search systems are based on supplying data with semantic annotations in order to improve search accuracy (precision and recall) on that data [Damjanovic, 2011; Mäkelä, 2012]. These semantic annotations are in fact metadata in the form of links to ontology concepts or individual descriptions in an appropriate semantic repository. The data the user is usually interested in are individual objects that may belong to particular classes and the domain knowledge is described primarily in terms of classes, their properties and relationships in the ontology. Thus if the properties point to objects, the search engine may ask the user to choose from the hierarchy of classes in the ontology the class of individual objects he is looking for and then apply the constructed keyword-based filters to the properties of the corresponding set of instances.

Various tools for semi-automatic and automatic semantic annotation have been used for the implementation of semantic search systems [Oliveira, 2013]. KIM (https://www.w3.org/2001/sw/wiki/KIM_Platform), ALIPR (http://meta-guide.com/ alipr-automatic-linguistic-indexing-of-pictures-realtime) and Onto-Mat-Annotizer (http://km.aifb.kit.edu/projects/annotation/Members/cobu/Annotation-Tool.2004-07-28.1138/view.html) are indicated as most popular semantic annotation platforms.

Usually semantic search engines provide tools allowing the user to formulate various kinds of complex queries as locating a group of individual objects of certain types connected by certain relationships. Properties and property relationships are used to traverse from a resource of interest to another. Sometimes the discovery of paths in the graph connecting objects is the expected search result. Another use case associated with the vision of semantic search concerns describing a problem and searching for its solution by reasoning on available knowledge bases [Rui, 2008].

Thus the use of appropriate types of metadata and heterogeneous semantic knowledge (thesauri, ontologies, semantic annotations) is the main characteristic of semantic search in its various forms and scenarios. The utilization of ontologies is considered as an approach to overcome the semantic heterogeneity of the searched repositories and datasets and therefore to achieve semantic interoperability between various information sources and search systems [Cao, 2004].

Some additional types of ontology applications as e.g. automatic reformulation of search queries to digital repositories containing semi-structured documents with incomplete and imprecise semantic annotations, ranking the reformulated queries, etc. characterize the new trends in semantic search [Mrabet, 2010].

3 Implementation of Semantic Search Engine in a Digital Library with Bulgarian Folk Songs

The digital library DjDL treasures a collection of over 1000 folk songs from the Thrace region of Bulgaria [Dzhidzhev, 2013]. The prototype of DjDL [Nisheva-Pavlova, 2011] was developed in 2010-2011 with the support of the Bulgarian National Science Fund and then a new version of DjDL that has some substantial features of a social semantic digital library was designed and implemented in 2014-2015 [Nisheva-Pavlova, 2015].

The folk songs preserved in the repository of DjDL are presented with their musical notations, lyrics and music (digitized versions of their authentic performances). The search engine of DjDL may be considered as a good example of semantic search engine.

The search engine of DjDL realizes two main types of search in the catalogue metadata and the lyrics of songs: keywords-based and semantic search. The semantic search tool provides a set of facilities for automatic reformulation (augmentation and refinement) of the queries for keywords-based search according to the available explicit domain knowledge.

Two forms of conceptual domain knowledge are maintained in DjDL – a subject ontology and a set of concept search patterns based on this ontology.

The subject ontology contains definitions of concepts, descriptions of their properties and several types of relationships between them, as well as a number of their representative instances. It describes an amount of knowledge in several domains, relevant to the content of Bulgarian folk songs: manner of life and family (professions, instruments, clothing, ties of relationship, feasts, traditions and rites, etc.), historic events, social phenomena and relationships, impressive natural phenomena.

The so-called concept search patterns are natural language-dependent patterns of typical stylistic or thematic constructs frequently appearing in the lyrics of Bulgarian folk art. They are defined and have been used as domain knowledge aimed at improving the precision and recall of the search engine.

The search engine realizes some additional functionalities enabling the user to combine the search and retrieval of documents kept in the repository of DjDL with a kind of sentiment analysis of their texts. The sentiment analysis tool uses the subject ontology as a source of knowledge about the emotional intensity of its concepts and computes rough estimates of the mood of songs.

The discussed search engine is developed as a client-server application built on the .NET Framework 4.5 and ASP.NET MVC 5. The tool used for its implementation is Microsoft Visual Studio Ultimate 2012 with additional packages for ASP.NET MVC 5. For JavaScript processing the jQuery library v. 1.10.2 has been used.

The implementation of the digital library system of DjDL is illustrated in Figure 1. It is based on Entity Framework 5 technology in combination with Code First. The current version of DjDL uses a local database (SqlLocalDB v. 11.0).

The class library RDFXMLClassLibrary was especially built for the purpose of automatic conversion of the original files with metadata and lyrics of songs (available in LaTeX format) to the RDF format in which they are processed by the search engine. LilyPond should be installed as an external software package on the server in order to generate files with standard musical notations (that are necessary for the adequate visualization of the search results) and MIDI files with melodies of songs from the original source files treasured in the repository of DjDL [Peycheva, 2010].

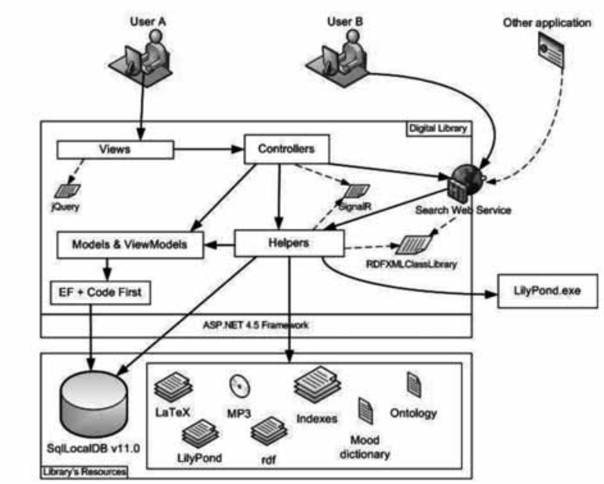


Figure 1. Implementation of the digital library system of DjDL [Nisheva-Pavlova, 2015]

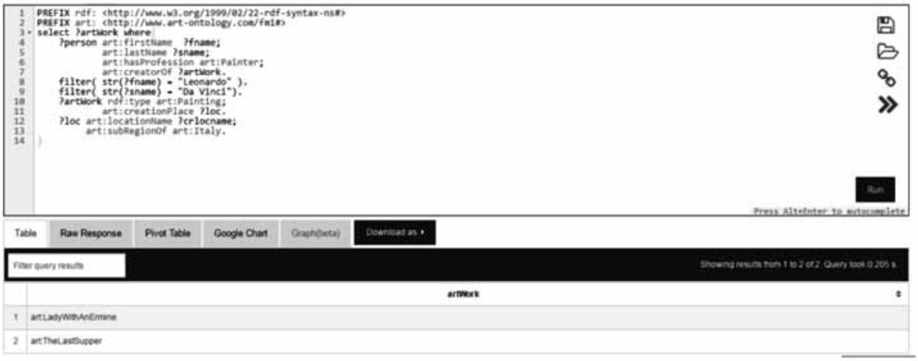
The subject ontology was built using the popular free ontology editor Protégé, version 4.3. Most concepts of this ontology are constructed as defined OWL 2 classes.

4 Experiments with Free Software for Advanced Semantic Technologies

A series of practical experiments with various free software tools supporting modern semantic technologies were performed in 2015/2016 academic year within the Master's degree course in Semantic Technologies at the Faculty of Mathematics and Informatics, Sofia University, in order to obtain some firsthand experience of their functionalities and user interfaces and to design and implement small projects aimed at development of semantic databases in chosen domains, formulation and execution of different types of queries for search and reasoning on these databases.

The most promising results were obtained with the free version of GraphDB (http://graphdb.ontotext.com/). GraphDB may be characterized as a semantic repository, compliant with W3C standards. It is a graph database system (also called a RDF triplestore) that supports the load, preservation, management, and querying digital content in the form of semantically enriched datasets in real time. GraphDB uses ontologies to perform automatic reasoning about data and to create new facts that are implied in data. Even its restricted free version is designed as an enterprise-grade semantic repository system, suitable for massive volumes of data.

We made more than 50 experiments with GraphDB Free and realized that it gives the developer a good set of facilities for rapid implementation of semantic databases and semantic search tools adequate for the requirements and expectations of most users. Within these experiments the size of the created and maintained semantic databases significantly changed. They were developed on the base of different sources – from relatively small RDF or OWL ontologies built especially for the occasion with the latest versions of Protégé ontology editor (http://protege.stanford.edu/) to the large DBpedia Ontology 2015 (http://wiki.dbpedia.org/services-resources/ontology) supplemented with specific DBpedia datasets.



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX art: <http://www.art-ontology.com/fm1#>
select ?artWork where
    ?person art:firstName  ?fname;
            art:lastName ?sname;
            art:hasProfession art:Painter;
            art:creatorOf ?artWork.
    filter( str(?fname) = "Leonardo" ).
    filter( str(?sname) = "Da Vinci").
    ?artWork rdf:type art:Painting;
            art:creationPlace ?loc.
    ?loc art:locationName ?crlocname;
         art:subRegionOf art:Italy.
}
```

Press Alt+Enter to autocomplete

Filter query results

Showing results from 1 to 2 of 2. Query took 0.205 s.

	artWork
1	art:LadyWithAnErmine
2	art:TheLastSupper

Figure 2. An example query for semantic search in GraphDB Free [Todorova, 2016]

It is worth to mention here at least two student projects aimed at creation and querying semantic databases. The first one [Todorova, 2016] is devoted to the world of pictorial and plastic arts and museums treasuring works of art. The semantic search is realized with the use of the "standard" OWL-Horst (Optimized) reasoning ruleset and a small training OWL 2 ontology. Figure 2 shows an example of semantic search query within this project.

The second project [Mutafchiev, 2016] is based on the utilization of a large ontology – the DBpedia ontology extended with some classes and relationships of the Friend-of-a-Friend ontology (http://xmlns.com/foaf/spec/). All experiments within this project demonstrated the good efficiency of GraphDB Free.

There is only one problem with GraphDB that needs a proper solution – the lack of interface module supporting the flexible and convenient (intuitive) construction of user queries for semantic search instead of the direct formulation of respective SPARQL queries.

5 Conclusion

Various free software environments that support modern semantic technologies can be used for the rapid development of flexible semantic search tools. Our experience in the design and implementation of a number of projects in the field of semantic search and information retrieval demonstrates that special attention should be paid to two particular issues concerning the user interfaces of semantic search engines: the convenient construction of complex user queries and the manageable presentation of search results.

Acknowledgements. The presented work has been partially funded by the Sofia University SRF within the “Methods for Automated Semantic Annotation of Heterogeneous Datasets” Project, Contract No. 15/2016.

References

Cao S. et al. (2004). Semantic Search among Heterogeneous Biological Databases Based on Gene Ontology. Acta Biochimica et Biophysica Sinica 36(5), 365–370.

Damjanovic V. et al. (2011). Semantic Enhancement: The Key to Massive and Heterogeneous Data Pools. In: Proceedings of the 20th International IEEE Electrotechnical and Computer Science Conference (Portoroz, Slovenia, 2011).

Dzhidzhev T. (2013). Folk Songs from Thrace. L. Peycheva, G. Grigorov, N. Kirov (Eds.), Sofia, Prof. Marin Drinov Academic Publishing House.

Guha R., McCool R., Miller E. (2003). Semantic Search. In: Proceedings of the 12th International World Wide Web Conference (Budapest, Hungary, 2003), 700-709.

Mäkelä E. (2005). Survey of Semantic Search Research. In: Proceedings of the Seminar on Knowledge Management on the Semantic Web, Department of Computer Science, University of Helsinki.

Mäkelä E., Hyvönen E., Ruotsalo. T. (2012). How to Deal with Massively Heterogeneous Cultural Heritage Data – Lessons Learned in CultureSampo. Semantic Web 3, 85–109.

Mrabet Y., Bennacer N., Pernelle N., Thiam M. (2010). Supporting Semantic Search on Heterogeneous Semi-structured Documents. In: B. Pernici (Ed.), CAiSE 2010. LNCS 6051, 224-229.

Mutafchiev D. (2016). Creating a Semantic Database and Execution of Search Queries in it. Course Project in Semantic Technologies, Sofia University, Faculty of Mathematics and Informatics.

Nisheva-Pavlova M., Pavlov P. (2011). Ontology-Based Search and Document Retrieval in a Digital Library with Folk Songs. Information Services and Use, ISSN 1875-8789, 31 (2011), 157-166.

Nisheva-Pavlova M., Shukerov D., Pavlov P. (2015). Design and Implementation of a Social Semantic Digital Library. Information Services and Use, ISSN 1875-8789, 35 (2015), 273-284.

Oliveira P., Rocha J. (2013). Semantic Annotation Tools Survey. In: Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 301-307.

Peycheva L., Kirov N., Nisheva-Pavlova M (2010). Information Technologies for Presentation of Bulgarian Folk Songs with Music, Notes and Text in a Digital Library. In: Proceedings of the Fourth International Conference on Information Systems and Grid Technologies (Sofia, May 28-29, 2010), St. Kliment Ohridski Uiversity Press, 218-224.

Rui H., Zhongzhi S. (2008). A New Approach to Heterogeneous Semantic Search on the Web. Journal of Computer Research and Development 45(8), 1338-1345.

Todorova G. (2016). Ontology of Art. Course Project in Semantic Technologies, Sofia University, Faculty of Mathematics and Informatics.

Wei W., Barnaghi P., Bargiela A. (2008). Search with Meanings: An Overview of Semantic Search Systems. International Journal of Communications of SIWN 3, 76-82.

Formal Specification of CAPECs in CSP

Vladimir Dimitrov,

Faculty of Mathematics and Informatics, University of Sofia, 5 James Bourchier Blvd.,
1164 Sofia, Bulgaria
cht@fmi.uni-sofia.bg

Abstract. Cyber-attacks are described in formatted text. There is no widely accepted formal notation for that purpose. This paper shows how CSP can be used for formal specification of CAPEC-47.

Keywords: cyber-attack, formalization, CSP.

1 What is CAPEC?

Some definitions from [1] follow below.

An **attack** is the use of an exploit(s) by an adversary to take advantage of a weakness(s) with the intent of achieving a negative technical impact(s). An attack includes the entire "Cyber Attack Lifecycle" reconnaissance, weaponize, deliver, exploit, control, execute, and maintain.

An **attack pattern** is an abstraction mechanism for helping describe how an attack against vulnerable cyber-enabled capabilities is executed. Each pattern defines a challenge that an adversary may face, provides a description of the common technique(s) used to meet the challenge, and presents recommended methods for mitigating an actual attack. Attack patterns help categorize attacks in a meaningful way in an effort to provide a coherent way of teaching designers and developers how their cyber-enabled capabilities may be attacked and how they can effectively defend them. Common Attack Pattern Enumeration and Classification (CAPEC™) provides a formal list of known attack patterns.

A **cyber-enabled capability** is any software enabled technology, irrespective of whether it be traditional information technology (IT), communications systems, industrial control systems, avionics, vehicle control systems, Internet of Things (IoT), or something that comes into existence next week. It also includes the interaction mechanisms such as Bluetooth, GPS, IR, Near Field Communication, USB, and other methods since these are all mechanisms for an attacker to influence the capability.

When considering attacks on cyber-enabled capabilities, we must address all aspects of those capabilities and how they are defined, designed, contracted for, produced, tested, acquired, delivered, maintained, serviced, and retired or disposed of. In addition, how they are used and interacted with, such as through

V. Dimitrov, V. Georgiev (Editors): ISGT'2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

physical buttons, switches, menu items, input fields, and keyboard/mouse input, must also be considered.

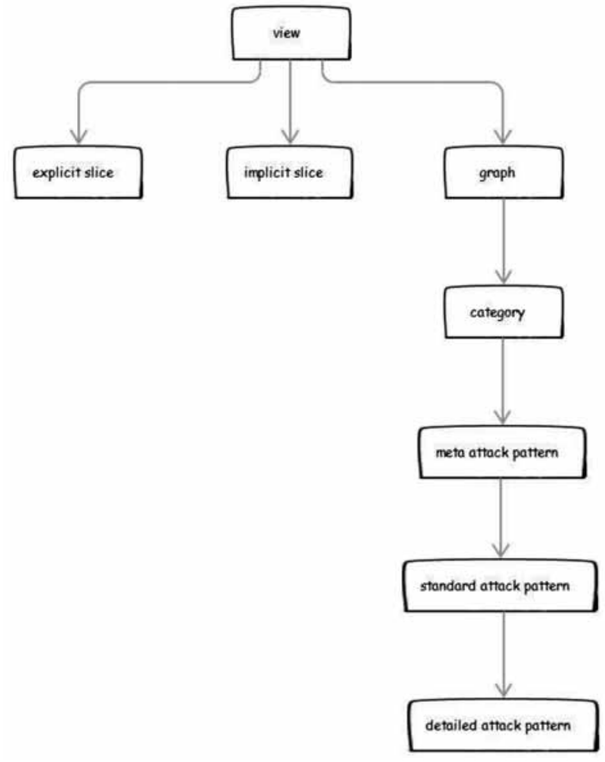


Fig. 1. CAPEC hierarchy.

A **view** in CAPEC represents a perspective with which one might look at the collection of attack patterns defined within CAPEC. There are three different types of views: graphs, explicit slices, and implicit slices.

A **graph** in CAPEC is a hierarchical representation of attack patterns based on a specific vantage point that a user may take. The hierarchy often starts with a category, followed by a meta-attack pattern/standard attack pattern, and ends with a detailed attack pattern.

An **explicit slice** in CAPEC is a subset of attack patterns that are related through some external factor. For example, a view may be used to represent mappings to external groupings like a Top-N list.

An **implicit slice** in CAPEC is a subset of attack patterns that are related through a specific attribute. For example, a slice may refer to all attack patterns in draft status, or all existing meta-attack patterns.

A **category** in CAPEC is a collection of attack patterns based on some common characteristic. More specifically, it is an aggregation of attack patterns based on effect/intent (as opposed to actions or mechanisms, such an aggregation would be a meta-attack pattern). An aggregation based on effect/intent is not an actionable attack and as such is not a pattern of attack behavior. Rather, it is a grouping of patterns based on some common criteria.

A **meta-level attack pattern** in CAPEC is a decidedly abstract characterization of a specific methodology or technique used in an attack. A meta-attack pattern is often void of a specific technology or implementation and is meant to provide an understanding of a high level approach. A meta level attack pattern is a generalization of related group of standard level attack patterns. Meta-level attack patterns are particularly useful for architecture and design level threat modeling exercises.

A **standard level attack pattern** in CAPEC is focused on a specific methodology or technique used in an attack. It is often seen as a singular piece of a fully executed attack. A standard attack pattern is meant to provide sufficient details to understand the specific technique and how it attempts to accomplish a desired goal. A standard level attack pattern is a specific type of a more abstract meta level attack pattern.

A **detailed level attack pattern** in CAPEC provides a low level of detail, typically leveraging a specific technique and targeting a specific technology, and expresses a complete execution flow. Detailed attack patterns are more specific than meta-attack patterns and standard attack patterns and often require a specific protection mechanism to mitigate actual attacks. A detailed level attack pattern often will leverage a number of different standard level attack patterns chained together to accomplish a goal.

The hierarchy of CAPEC is given in Fig. 1. It only abstraction levels based.

As an example CAPEC-47 from [1] is given below:

“

CAPEC-47: Buffer Overflow via Parameter Expansion

Attack Pattern ID: 47

Abstraction: Detailed

Status: Draft

Completeness: Complete

Presentation Filter: Basic Complete

Summary

In this attack, the target software is given input that the attacker knows will be modified and expanded in size during processing. This attack relies on the target software failing to anticipate that the expanded data may exceed some internal limit, thereby creating a buffer overflow.

Attack Prerequisites

- The program expands one of the parameters passed to a function with input controlled by the user, but a later function making use of the expanded parameter erroneously considers the original, not the expanded size of the parameter.

- The expanded parameter is used in the context where buffer overflow may become possible due to the incorrect understanding of the parameter size (i.e. thinking that it is smaller than it really is).

Solutions and Mitigations

Ensure that when parameter expansion happens in the code that the assumptions used to determine the resulting size of the parameter are accurate and that the new size of the parameter is visible to the whole system

Related Attack Patterns

Nature	**Type**	**ID**	**Name**
ChildOf	S	100	Overflow Buffers

“

Every attack in CAPEC has a name and ID. In this example ID is 47. Abstraction level is “detailed” that means “detailed level attack pattern” – see above the term definition. The information provided in the other sections of this descriptions is simple and clear – not very much detailed.

One more thing is the context of CAPEC-47. It is a leaf of hierarchy shown in Fig. 2.

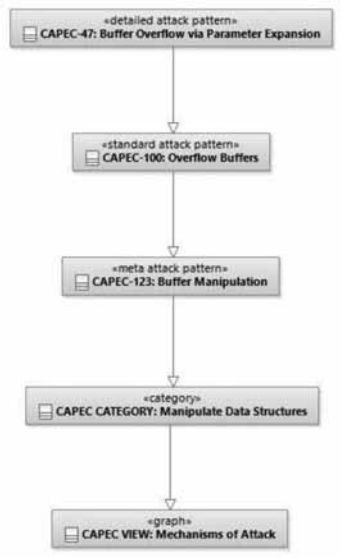

«detailed attack pattern»
CAPEC-47: Buffer Overflow via Parameter Expansion
«standard attack pattern»
CAPEC-100: Overflow Buffers
«meta attack pattern»
CAPEC-123: Buffer Manipulation
«category»
CAPEC CATEGORY: Manipulate Data Structures
«graph»
CAPEC VIEW: Mechanisms of Attack

Fig. 2. CAPEC-47 context.

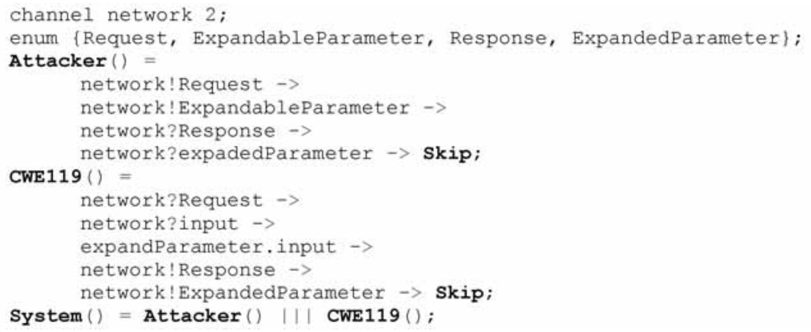
The context of CAPEC-47 is simplified, because it is very complex at the higher levels.

CAPEC-47 in the context of CVEs and CWEs have to be analyzed. Let’s start with Heartbleed vulnerability. First at all, the vulnerability is detected. Searching CVEs it is CVE-2014-0160 (Heartbleed). In “Technical Details” section mentions only CWE-119 is mentioned as only applicable weakness. In “Related Attack Patterns” section of CWE-119 among other attacks is mentioned CAPEC-47, because the payload is expanded by the server. Therefore vulnerability CVE-2014-0160 (Heartbleed) is a result of successful attack CAPEC-47 of the weakness CWE-119.

2 CAPEC-47 Formalization

Candidate tools for formal specification of attacks are CSP and UML. Communicating Sequential Processes (CSP) [2] is well-designed algebra. It is useful for distributed application specification and verification. It is behavioral in nature and can be used for CAPEC specification (attacks). The problem with it is the limited set of tools supporting the notation – mainly at academia.

For verification purposes Pat tool [3] is used. The CAPEC-47 specified in CSP is:



```
channel network 2;
enum {Request, ExpandableParameter, Response, ExpandedParameter};
Attacker() =
      network!Request ->
      network!ExpandableParameter ->
      network?Response ->
      network?expadedParameter -> Skip;
CWE119() =
      network?Request ->
      network?input ->
      expandParameter.input ->
      network!Response ->
      network!ExpandedParameter -> Skip;
System() = Attacker() ||| CWE119();
```

First, a channel named “network” is defined. This channel is buffered and can accept no more than 2 messages. This way of simulating the network with a channel is somehow abstract – the channel reliably deliver the messages.

Messages are defined as enumerations: Request, ExpandableParameter, Response, and ExpandedParameter.

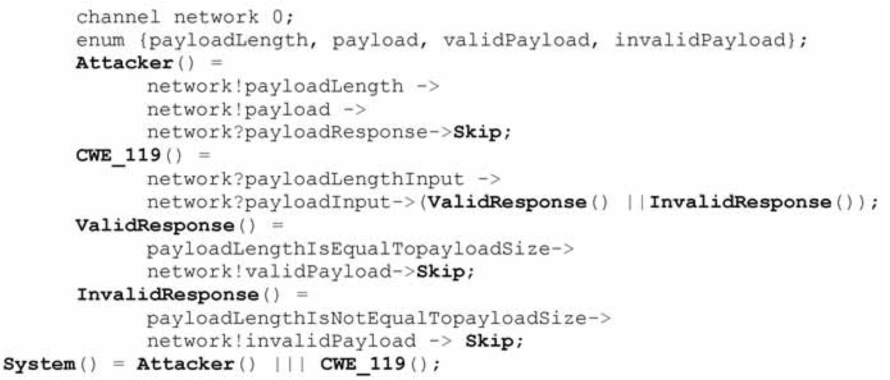
There are 2 processes: Attacker and CWE119. The first one performs the attack and the second one exposes and realize the vulnerability Heartbleed. These processes are executed in parallel – they build the main process System. Their synchronization is based on reads and writes on the channel (network).

The attacker sends a request to the vulnerable process. This request contains an expandable parameter, i.e. this parameter contains a text message with its length. Then starts to wait for a response with an expanded parameter from CWE119.

The CWE119 receive a request from the attacker. The request contains an expandable parameter. The trick is that length of the message is bigger than the actual message size. When the vulnerable process receive the text message, it create a buffer for the message using the received length but not its actual size. The message is placed in the beginning of the buffer, but the remaining part of the buffer stays unchanged containing information that is possible to include private keys. The whole buffer is send back to the attacker – not only the exchanged text message. Now the attacker can investigate the received information for sensitive data.

Note that CAPEC-47 is described in very generic manner at a very high level that is why the CSP specification is very simple and in that case not very useful.

The following is another attempt to specify the Heartbleed bug in pure CSP:



```
        channel network 0;
        enum {payloadLength, payload, validPayload, invalidPayload};
        Attacker() =
              network!payloadLength ->
              network!payload ->
              network?payloadResponse->Skip;
        CWE_119() =
              network?payloadLengthInput ->
              network?payloadInput->(ValidResponse() ||InvalidResponse());
        ValidResponse() =
              payloadLengthIsEqualTopayloadSize->
              network!validPayload->Skip;
        InvalidResponse() =
              payloadLengthIsNotEqualTopayloadSize->
              network!invalidPayload -> Skip;
System() = Attacker() ||| CWE_119();
```

In this example the network is specified as unbuffered facility. This means that when a process sends a message, it remains blocked until the other process do not accept the message.

There are four kinds of messages payloadLength, payload, validPayload and invalidPayload.

The system consists of two interleaving processes: Attacker and CWE_119.

The Attacker sends in reality one message, but it is divided in two parts payloadLength and payload to focus on the fact that the payloadLength is not equal to the real payload. Then the Attacker wait for the response.

The process CWE_119 accepts payloadLengthInput and the payloadLength. Then there are two events (cases) when the payload length is equal or not to payload size. In the first case ValidResponse process is executed returning a valid payload that is the case of the normal execution of CWE_119 weak process. The other case when the accepted payload length is not equal to the real payload length and then the process InvalidResponse is executed returning invalidPayload.

In this example CWE-119 is used intensively. The specification can be reworked in more details in PAT dialect, sent an array, check for its size, etc.

There is a need also to add an assert that the network!invalidPayload state is reached, but it will ruin the clean specification.

Even more detailed specification of Heartbleed can be given in CSP, but so detailed has to be – at C language level. So, why not such attacks to be described directly in their programming language?

3 Conclusion

In this example, the CSP specification is very generic following the CAPEC-47 description. Anyway, the specification can be in more details if knowledge from weaknesses is used. Finally, this specification can be detailed at the programming language level (in that case C) using CSP.

Is useful such a specification? CSP specification of an attack is behavior specification. This can be used in tools for attack detection. Very detailed specification contains many internal events that are not available for attack detection tool. Some external events are in their nature checks on the message as the event payloadLengthIsNotEqualTopayloadSize. This means that some standard for external events must be developed if we want these specification to be inputs for attack detection tools.

Actually, first vulnerability is detected, then it is described as weakness(es) and attack exploiting this weakness(es). Attacks without weaknesses are very high level and unusable. They must be investigated and specified together with the weaknesses. Pragmatically, vulnerability must be the starting point for specification, then the attack with the weaknesses must be specified.

It is clear, that further research on the specification of attacks must be done.

References

1. Common Attack Pattern Enumeration and Classification (CAPEC™), http://capec.mitre.org
2. Hoare, C. A. R. Communicating Sequential Processes. Prentice Hall International. ISBN 0-13-153271-5 (2004) [1985].
3. Sun, Jun; Liu, Yang; Dong, Jin Song. Model Checking CSP Revisited: Introducing a Process Analysis Toolkit. Proceedings of the Third International Symposium on Leveraging Applications of Formal Methods, Verification and Validation (ISoLA 2008). Communications in Computer and Information Science. Springer. pp. 307–322. Retrieved 2009-01-15 (2008).

Roboethics: questions related to the implementation of robots

Ioannis Patias

Faculty of Mathematics and Informatics
University of Sofia St.Kliment Ohridski“
5 James Bourchier blvd., 1164, Sofia, Bulgaria
ioannis.patias@gmail.com

Abstract. This paper presents the issues (social and ethical) raised by the wider spread of commercialized application of robots to our daily life. The applications of robots to society are expected to be wider and wider, and robotics is going to trigger social and economic changes. This opens new social and ethical questions for which the designers, and the constructors must now be familiar and prepared. Starting from an historical review of the understandings of robots, the paper summarizes the recent developments, and the relationship between theory and practice in legal, ethical, financial, economical etc aspects. The questions presented should be well understood by each area specialists, engineers, designers, and constructors. However, given the subject is complex the conclusion is that those ethical problems need deeper analysis and require the combined efforts of many professionals.

Keywords: robots, roboethics, machine ethics.

1 Introduction

Roboethics is a short expression for ethics of robotics. It is often used in the sense that it is concerned with the behavior of humans, how humans design, construct, use and treat robots and other artificially intelligent beings, whereas machine ethics is concerned with the behavior of robots themselves, whether or not they are considered artificial moral agents (AMAs) [1]. While the issue is as old as the word robot, the short word roboethics was put forward in 2001/2002, and publicly discussed in 2004 during the First International Symposium on Roboethics by roboticist Gianmarco Veruggio [2].

Even in mythology the idea of robots is introduced from ancient times mankind. Ancient civilizations and religions have similar views. For example:

- Hephaestus (Ἥφαιστος) in Greek, and
- Vulcan (Volcānus), in Roman mythology.

The perception of robots in mythology is more associated with religion than with technology. Hephaestus used to build machines from metal to work for him.

V. Dimitrov, V. Georgiev (Editors): ISGT'2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

This included statues that went to and from Mount Olympus. Greek myths and Homeric poems describe how Hephaestus gave them special power to produce movement. So somehow humanity then considered the automated machine was god or a statue which got life from god, and thus was the continuation of the god himself.

2 Perception of robots

The first steps in the direction of perception of robots, as a product of technology, are made by Isaac Asimov [3, 4]. Asimov remains in history with the introduction in 1942 of three laws of robotics. Laws must be built into every robot and thus to ensure its safe use. The Three Laws of Robotics (often shortened to The Three Laws or Three Laws, also known as Asimov's Laws) are a set of rules devised by the science fiction author Isaac Asimov. The rules were introduced in his short story "Runaround", although they had been foreshadowed in a few earlier stories. The Three Laws, quoted as being from the "Handbook of Robotics, 56th Edition, 2058 A.D.", are:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

In many situations Asimov's laws get criticized [5]. Having in mind the position that Asimov wrote his laws focused on using them for his fiction stories, in one of Asimov's stories, robots were made to follow the laws. But again they were given a certain meaning of "human". In real-world situations of ethnic cleansing campaigns for instance, we have robots, which only recognize people of a certain group as "human". And in such case even following the laws, they still may carry out genocide. But again this puts under question how robots are being used in our real world. Human arms a drone with a missile or puts a machine gun on a robotic system, and still asks not to cause humans to come to harm. It is also similar when we build robots that take orders from any human. We do not really want any human to be able to order a robot.

We build robots, which can be sent out on dangerous missions and be destroyed (killed) and we consider this very rationale to using them. But, what will happen when we add to them a sense of "existence"? The survival instinct in this case goes against that rationale. What we see here is that for robotic systems coming from the military, the robots need to follow very opposite laws of those defined by Asimov. That means we define robots which first can kill, second they don’t take orders from any human, and they have no survival instinct to care about their own existence.

What we see here is that concerning robots and ethics is not whether we can define laws like Asimov’s laws, and thus make machines that are moral, but, how to influence the ethics of the people building the machines. We need a code of ethics in the robotics defining what gets built, and who can use such sophisticated systems rather than blaming the machines.

3 The modern reality

In modern reality deontology, i.e. the formation of a set of rules of conduct, is not suitable neither for human nor for roboethics and machine ethics.

What can be said that there is a practical and applied focus are the specific legislative initiatives. The first specific law that aims to define a particular problem and give specific direction to resolve it is a federal “Law on the modernization and reform of 2012.” (FAA Modernization and Reform Act of 2012) [6]. It is specifically related to the incorporation of unmanned aircraft in systems administration, plans and policies of the Federal Aviation Administration.

In Europe, things are also very dynamic. The European Aviation Safety Agency (EASA) published on 18 December 2015 a formal Technical Opinion on the operation of drones [7]. This opinion lays down the foundation for all future work for the development of rules, guidance material, as wells as, safety promotion to ensure unmanned aircraft are operated safely and their impact on the safety of the aviation system is minimized. The opinion includes 27 concrete proposals for a regulatory framework for low risk operations of all unmanned aircraft irrespective of their mass. The proposals are operations centric, focusing on how the drones will be used rather than their physical characteristics. It establishes 3 categories of operation: ‘Open’, ‘Specific’ and ‘Certified’ with different safety requirements for each, proportionate to the risk.

But except the example of aircraft, many are actively working also in the direction of establishing regulations for autonomous vehicles. Since the end of 2015, only in California were issued 11 permits for testing of autonomous vehicles in real conditions [8]. These companies are directly interested in defining the rules in order to begin mass production, and this is not an easy process to model [9] [10].

4 The future raises questions

As a whole due to the relatively small penetration of robots so far (relatively small number of real, widespread, commercialized applications) in the near future we will see a lot of developments in this area. For example, if we try to look at the areas of developing applications, grouping them on the interface we see:

- Human-softbot integration: AI for information & communication.

distributed computing for human assistance
intelligent agents for information and communication management
edutainment AI

- Human-robot, non-invasive integration: Autonomous robotics.

personal and assistance robotics edutainment robotics
warfare application of robotics

- Physical, invasive integration: Bionics.

prosthesis
enhancement of human sensorimotor capabilities ICT implants

But the future raises questions, which we can group based on the expertise which is required for their solution as:

- engineering

Is it possible to create a machine that can distinguish a gun of an ice cream or fully understand human speech, often heavily based on context?

How we balance the need to preserve the robots from amok with the need to protect from hacking or capture?

What would be the impact of robotic industry on the environment?

What are the trade-offs between robotic solutions and safety, for example:

  - soft mechanisms, soft robotic limbs or organs?
  - and what about when using weapons?
  - to use their help only in certain situations, such as when it is assumed that all humans around are enemy targets?

- end-user related

How safe should be the robots before introducing them into the market or society?

Is there a risk of emotional attachment to robots?

Is there something essential in human communication and relationships that robots can not replace?

- financial and economical

What is the estimated economic impact of robotics? how to calculate the expected costs and benefits?

There are some jobs too important or too dangerous to be undertaken by machines? But what do we do with workers displaced by robots?

How to reduce the risks of a society dependent on robots? if these robots become unusable or damaged, for example by electromagnetic pulse, or a virus?

- legal

Whether we have the moral right to pass on our responsibility to our children and adults to machines? whether they are a substitute for human interaction?

Robot interaction and companionship to be used for other purposes, such as pets or sexual partners?

At what point a robot becomes “human” in terms of rights and responsibilities?

Do we need to have a different legal status for robots, different than for ordinary people?

5 Conclusions

In brief was presented the perception of robots of humanity and how it changed over time. It went from abstract perception of robots as an expression of God’s power and capabilities (Hephaestus) in science fiction, and the introduction of common sense (Isaac Asimov), to the initial attempts to specifics, and the first legal formalization (FAA Modernization and Reform Act of 2012) and subsequent initiatives. Also, they were grouped by interface the prospects of developing a wide range of robotics applications.

In conclusion, the topic is very broad; the ethical questions that arise are many and require the combined efforts of many professionals.

References

1. Peter M. Asaro, What Should We Want From a Robot Ethic?, International Review of Information Ethics, www.i-r-i-e.net, ISSN 1614-1687, Vol. 6 (12/2006)
2. Gianmarco Veruggio, Symposium Chair, ROBOETHICS ROADMAP (http://www.roboethics.org), The ethics, social, humanitarian and ecological aspects of Robotics, 2004
3. Asimov, I., I Robot, Doubleday, 1950
4. Asimov, I, Runaround, Astounding Science Fiction, March 1942. Republished in Robot Visions by Isaac Asimov, Penguin, 1991
5. Barthelmess U, Furbach U, Do we need Asimov’s Laws?, http://www.cornell.edu, http://arxiv.org/abs/1405.0961, 2014
6. FAA Modernization and Reform Act (P.L. 112-095) Reports and Plans, https://www.faa.gov/about/plans_reports/modernization/, 2012
7. European Aviation Safety Agency (EASA), https://www.easa.europa.eu/document-library/opinions/opinion-technical-nature, 2015

8. California Department of Motor Vehicles, https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/testing, 2016
9. Evgeniy Krastev, Maria Semerdjieva,"Business Process Model Based on Business Rules", Proceedings of the 8th Int. Conference Information Systems & GRID Technologies 2014
10. Evgeniy Krastev, Kristiyan Shahinyan, Computer Assisted Quality Assessment of a Set of Business Process Models, Proceedings of the 9th IEEE European Modelling Symposium of Mathematical Modelling and Computer Simulation, IEEE Computer Society, 2015, pp.180-186, doi:10.1109/EMS.2015.36

Modelling of GIS Based Visual System for Local Educational Institutions' Stakeholders

Snezana Savoska[1], Valmir Muaremi[1], Andrijana Bocevska[1], Blagoj Ristevski[1] and Zoran Kotevski[1]
[1]Faculty of Information and Communication Technologies
University „St.Kliment Ohridski" – Bitola,7000,ul. Partizanska bb, Macedonia,
snezana.savoska@fikt.edu.mk, valmir.m070@gmail.com, andrijana.bocevska@fikt.edu.mk,
blagoj.ristevski@fikt.edu.mk, zoran.kotevski@fikt.edu.mk

Abstract. Application and implementation of the newest technologies represent significant obligation of government particularly in primary education, as starting point for further level of education in order to create the habits of using them. The current decision making systems in education are mostly connected to the newest technology trends as visual data analysis and GIS systems, which demands training of analytics staff. Many efforts and resources are required firstly to provide training for educators to use the newest information technology and then to create system which can help to monitor data about all level of education and all involved stakeholders. One of the desired IT challenges is the usage of visual data representation with GIS systems for stakeholders in education. We start with modelling complex visual GIS based system intended for stakeholders of dispersed schools through rural and others regions. It has to consider many factors: data and legal obligations set from the state' legal framework, needed transport and costs that are on charges on the local government. Additionally, a multilanguage support has to be implemented in the local institutions because of the legal frame in selected areas. The model differs from the used central data model for primary education and it should deal with missing data caused from schools' dislocations, insufficient number of pupils in some schools, insufficient teaching staff and unavailability of local schools as a result of missing road infrastructure and similar issues. Leading by these findings, we propose modelling of visual system based on GIS, intended for local primary educational institutions equipped with tools for business intelligence and visual data analysis which will help to solve some stakeholder's problems and demands in education.

Keywords: GIS, Visual Data Representation, System Modelling, GIS Based Data Visualization, Primary Education.

1. Introduction

The primary education is one of the most significant pillars of the state and for this reason the newest technologies have to be accepted and implemented firstly by the stakeholders in this educational sector. The actual decision making systems are connected mostly to visual data representation and usage of GIS systems and they have to be implemented in government sectors, especially in education. The purpose of application of these systems is to create competent staff, to train people for education in order to be able to follow actual and emerging IT trends. Many efforts and resources are required firstly to provide educators to be trained to use the newest information technology and then to create a suitable system that can help to monitor data about all level of education and all involved stakeholder, such as government institutions, municipality institutions, schools, parents, pupils, students etc.).

The key segment of education is the primary education and this is the first place where IT should be used in practical and efficient manner. One of the desired IT challenges is the usage of visual representation of data using GIS systems for stakeholders in education. We start with modelling complex visual GIS based system intended to stakeholders of dispersed schools through rural and others regions. By modelling many factors should be considered such as data and legal obligations set by the state' legal framework, the needed transport and costs which are on charges on the local government as well as multilanguage support which should be implemented in the local subsystems. The model should be different from the used central data model for primary education and should handle missing data caused from schools' dislocations, insufficient number of pupils in some schools, insufficient teaching staff and unavailability of local schools as a result of missing road infrastructure and similar issues.

Leading by these findings, we propose modelling of visual system based on GIS, intended to local primary educational stakeholders with usage of tools for business intelligence and visual data analysis that can help to resolve some demands for information about some ongoing education problems.

The rest of the paper is organized as follows. In the second section, related works are explained. In the third section, the prerequisites for the planned model was considered as well as the undertaken activities. The fourth section underlines the used platform with highlighting model, used software tool and implementation processes whose results are explained in the next section with focus of the gain for stakeholders of the primary educational institutions. Finally, some conclusion are drawn and future researches are proposed.

2. Related works

Many efforts are made for using GIS systems in all levels of education [6] [16] [17]. Some of them are related to usage of GIS for teaching with GIS, while the other on teaching about GIS. Some papers are trying to distinguish GIS at three level: GI Systems, GI Science and GI Studies [12] and analyze GIS involving in decision making processes, its penetration in human processing which rewrites our spatial practices, knowledge and decision making and conclude that is the quiet revolution which change customers behavior and human cognition [5, 21].

Forer and Urwin (1996) considered GIS progress in education [12] and stated that in the period from 1989 to 1996 to meet students' needs for cheap software, many raster systems have been designed to run on very basic hardware platform and they provided GIS education with many excellent tutorials systems ready to run laboratory materials, learning distance units and analog videos materials, mostly for tertiary education institutions [2]. At the end of this period, the conceptual model of end-use created by these efforts, produce many specific projects for defining institutional requirements with fairly rigid restrictions. The result was embedded GIS in various structures to facilitate entirely new learning outcomes particularly in the area of market, using technological forces [16]. They also define the factors for many emerging demands as shifting of GIS on marketing data usage, demands for user-friendly GIS, distributed and web-oriented GIS as well as declining overall costs for GIS and wider computer usage [16]. The flexibility of GIS also enables data oriented and software tools connected with databases, created from different ethical, social and economic reasons [16].

Usually GIS is used in high degree of education, in the secondary schools in some places and rarely in lower education institutions, mainly for learning geography courses [3, 10]. Despite the barriers and obstacles from the lack of time for teachers to learn GIS and use it for their class activities, unwillingness of teachers to learn and use new technologies and insufficient representation of GIS in the subjects 'syllabuses [3], some projects are trying to introduce GIS systems in education in different levels and curricula, especially in the secondary schools [14]. For example, GISAS, a project founded from MINERVA Action of the EU Commission, was water quality analysis project with usage of GIS around seven schools in seven EU countries within a web-based learning environment [16].

Also, the application of GIS in applied learning increases during the time [3]. We can mention the book Mapping Our World: GIS lessons for education published from ESRI (with ArcView 3.x for middle and high schools geography) providing another approaches of usage GIS in education. The next example is the project of application of GIS in secondary schools for geography curriculum in Turkey. That resulted with educated teachers and developed GIS-based application and its implementation in the schools that have the necessary hardware and software resources [3]. Many other projects was running at this time, which are more or less successfully.

Bernarz and Ludwig [2] had stated that diffusion of GIS in education has to be first implemented in higher education by GIS specialist with possibility of innovations in education [2, 20]. But, they stated that this process usually takes four stages: awareness, understandings, guided practice and implementation. Also, usage of GIS works well for solving real-state problems with constructivists' theories. Many examples are described for usage of GIS in solving local ecological problems with data collection as well as identifying business opportunities depending on locations which confirm that theory [16]. They highlighted the power of GIS in usage for visualization and interpretation of geographic information, as well as the reasons for incorporating GIS technologies in geography and other modern environmental science as a valuable tool for analysis and problem solving. They indicate the needs for increasing teachers' understand special thinking, understanding how GIS tools can construct spatial understanding and human cognition and learning. All these facts aim to reform educational curricula in all levels [16].

Other researchers [12] consider and interpret GIS as GISy – Systems, GISc – Science and GISt – Studies. GISy are focused on technology for the acquisition and management of spatial information, GISc are focused on conceptual issues of representing spatial data with deeper understanding of the meanings and creative analysis, while GISt are more specialized on social, legal and ethical issues with greater importance and complexity. They figured out increasing of market of cheap software, raster systems designed to run on very basic hardware which provided GIS education with many great tutorials and video materials. Many factors shifted GIS in using with marketing data, user-friendly software with distributed data, wider computer adoption and decreasing the overall costs. Since that period, GIS is widely used in education, especially in geographic curricula as well as strategic tools for marketing data analysis, agriculture economy with exploratory spatial analysis [13], human capital knowledge analysis [11], health care analysis [18], government issues toward e-commerce and e-government support [11, 4, 12].

Some projects was focused to bring symbiosis of GIS with others technologies as remote sensing activities, providing information for wider communities, academic research and business using distributed, local and hybrid models [1]. Some of them moved into cloud with defined semantic specification for spatial data infrastructure, provided interoperability and framework for web application based on open geospatial standards [7]. These steps already provide frame for geospatial web services, cloud GIS and semantic web for GIS, allowing data to be shared and reused, providing interoperability and spatial data infrastructure (SDI) and GML (Geography Markup Language) as XML based language [15].

Earth Observation (EO) data become available on the web from government agencies as results of increasing technological capacities [9, 7]. Next steps are connected with creating standards as OGC Web Services (OWS) enabling Cloud services quality requirements set by INSPIRE directives (EU directives) [8] in order to provide compatible Geospatial Open Source Software standards and semantic interoperability with tracking changes possibilities [21]. These facts lead to numerous projects that support these standards and produce many software tools running according to set standards [7]. In the latest ESRI User Conference held 27.6.2016, ESRI president highlighted that GIS enable a smarter World integrating Internet of Things, enabling framework and processes for enabling smart World, integrating and managing innovations, connecting peoples, things, processes and data about them with Smart GIS applications. They assume that ESRI has 3.2 million users, 94 billion maps views and 6.2 million items as datasets which have to enable SMART Web GIS as "nervous" system of the Earth [5]. This concept demand a huge engagement of all educational levels to direct teaching GIS at schools, as trainings and lifelong learning [5]. It is a part of the concept of smart leaving, smart cities with all infrastructure using GIS in each part of leaving (Fig.1) [19].

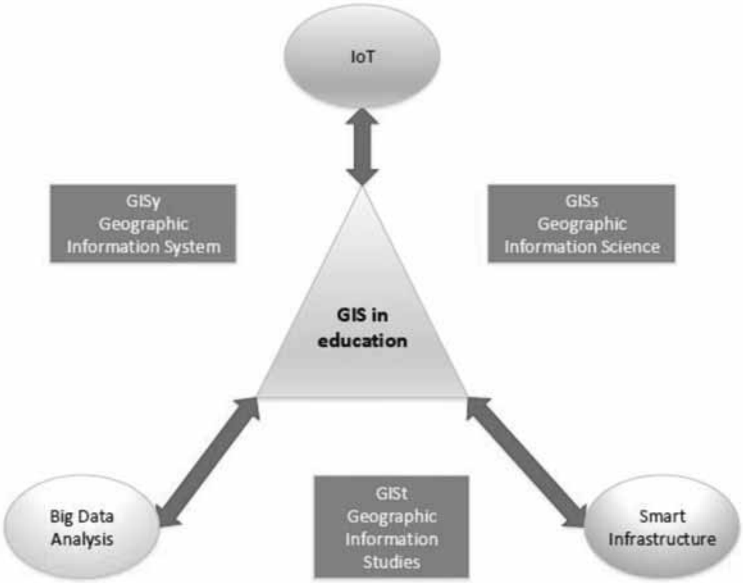


Figure 1 – Smart use of GIS-based systems

3. The prerequisites for the planed model

The research for gaining information for prerequisites needed for planned model was made in April 2016 in the rural region of Prespa, Resen municipality in Macedonia. The survey was intended for stakeholders in primary education as pupils, their parents, teachers, employees in Ministry of education and science, local government employees, and transporters. In some rural areas there is no local primary or secondary schools and pupils have to travel from villages to the nearest primary schools as well as secondary schools. Also, the population migration produces decreasing of the number of pupils especially in the rural regions as the selected one. Many factors' changes were made depending on migration activities in these regions, as available teaching staff, decreasing the numbers of classes, and open schools. Sometimes, government, local government, ministry of education, teachers and pupils don't have information about the real state with the schools, transport activities, classes, numbers of pupils in each class, real situation with the pupils, teachers and many other issues that influent the quality of education and total costs. Many activities connected with model creating are in accordance with previous related works.

For this purpose, the survey for teachers, pupils, transporters and other stakeholders in education was made. The collected data will be helpful to make a data selection, needed for creating a model that is GIS based visual system for local education institution stakeholders (VS4LEIS). The survey has taken in consideration the information demands of all stakeholders and it facilitated creating firstly conceptual model of GIS-based VS4LEIS (Fig. 2). According to the conceptual model, the needed UML models can be created as well as the metadata about database for creating this GIS based VS4LEIS.

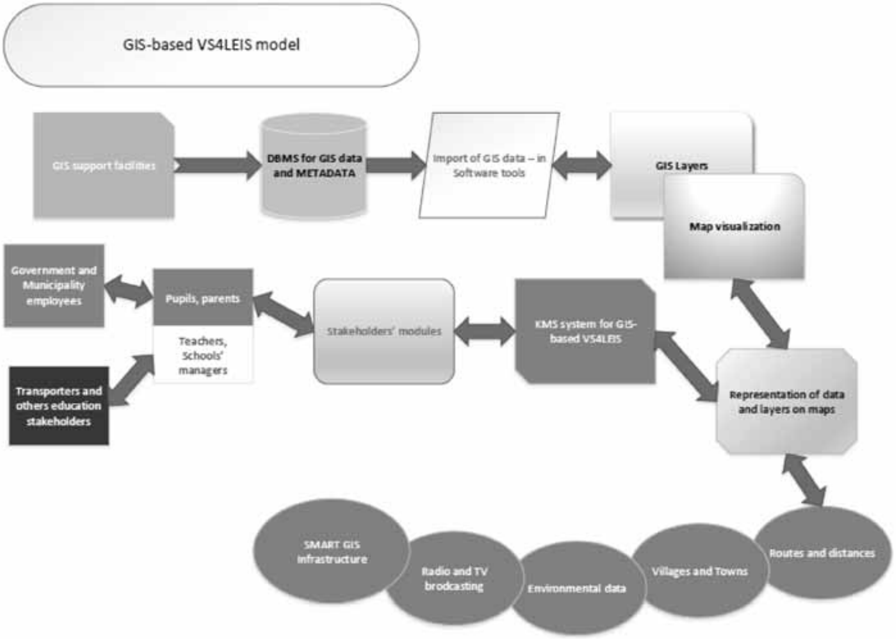
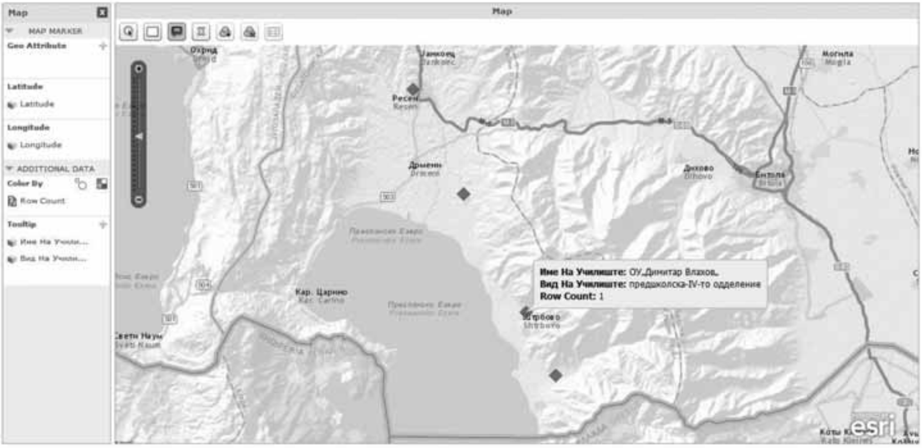

GIS-based VS4LEIS model
GIS support facilities
DBMS for GIS data and METADATA
Import of GIS data – in Software tools
GIS Layers
Map visualization
Government and Municipality employees
Pupils, parents
Teachers, Schools' managers
Transporters and others education stakeholders
Stakeholders' modules
KMS system for GIS-based VS4LEIS
Representation of data and layers on maps
SMART GIS infrastructure
Radio and TV brodcasting
Environmental data
Villages and Towns
Routes and distances

Figure 2 – Proposed model of GIS-based VS4LEIS

4. Platform and used software tool

The chosen platform should be one of the free platform for GIS and data visualization as a result of budget limitations of this project. Micro Strategy Analytics (MSA) is chosen for testing model implementation based on its good reputation on the software market tools and its good performances. After software installation, we analyzed compatible and suitable data models for this tool and model database structure. Next step in the modelling was to prepare data in suitable formats and to import these data into MSA tool. We have to choose the most effective data visualization on the GIS screens in couple of layers and to create knowledge management system (KMS) for GIS data based on the proposed VS4LEIS model. The obtained visualization of the test data can produce many layers of visual representation and one of them is presented on Fig.3. After these steps, we create the stakeholders screens and human-computer interface (HCI) for separated groups of stakeholders such as pupils and parents, teachers, schools' managers, government institutions, transporters and municipality's personals.



Име На Училиште: ОУ„Димитар Влахов„
Вид На Училиште: предшколска-IV-то одделение
Row Count: 1

Figure 3. GIS-based VS4LEIS data representation layer, created for specific example of regional data for Prespa-region

4.1. The explanation of the proposed GIS-based VS4LEIS model

The first part of the model defines GIS support facilities which prepare the base for satisfying users' demands. System analysis made by the authors has shown that the range of needed data for this GIS-based VS4LEIS is wide, object-oriented. That means that many objects have to be presented on the GIS-based system for separated groups of stakeholders. For this reason, data should be prepared on the separated tables and each table creates separate layer of data representation. Each stakeholder can have entity-relational matrix (ERM) for predefined rights and obligations for the system, usage and updating data and representation layers, managed by the system administrator.

All needed data should be saved in KMS system for GIS-based VS4LEIS together with ERM and some representations of the data are saved as historical data changes in an appropriate form of data warehouse.

The proposed model is dynamic, it changes in time by the authorized persons with defined HCI possibilities, with dynamic objects and data for the objects as their features and dynamic changing characteristics. In the model, all needed objects are taken into consideration, as routes and distances, villages and cities, environmental data, schools, teachers and pupils, students, parents, radio and TV broadcastings, teaching languages (Macedonian, Albanian and Turkish in one sentence – all smart city infrastructures' objects needed for smart living.

4.2. GIS-based VS4LEIS model implementation

Implementation of the proposed model is made on selected Prespa region in Macedonia that is represented as rural model of GIS-based VS4LEIS. This region possess all features of Macedonian rural regions and for this reason as a very representative region, we collected data from this region with purpose to test the proposed GIS-based VS4LEIS model. Data was collected according to proposed data model and stored in Excel files, which can be exported in many different database management systems (DBMS). According to the data in each spreadsheet, separate visualization layer can be created. Each table and layers is stored in KMS module for GIS-based VS4LEIS with version correlated with data and time of creation. This KMS module should enable development suitability of the model in project with saving historical data changes and their visual representations. Some data representation layers for collected data for selected Prespa-region are shown on Fig. 4a, 4b and 4c.

5. Results and discussion

The proposed GIS-based VS4LEIS model was created with purpose of obtaining needed data for local educational stakeholders. The created model demands a wide society engaging for collecting and updating data needed for these groups of selected people. The model is object-oriented and demands defining objects and stakeholders which will be represented in dynamic way, obtaining a sustainable model with historical data and ability for data analysis in time, in period, by each object or group of objects, groups and for the whole system. For this reason, as a pilot project, we implement the model for region of Prespa, with small numbers of objects and their features as data in the database tables. The proposed data model suggests usage of codes in many levels as subgroups for each region in order to obtain data analysis grouped by the objects of interest.

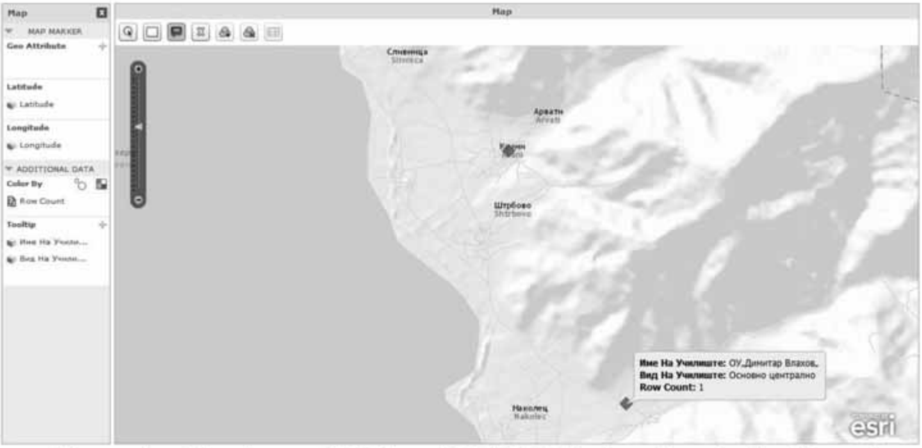
5.1. Gaining for stakeholders

In general, there are a couple of stakeholders' groups with similar interests which have to obtain benefits from the proposed GIS-based VS4LEIS. First group are schools' personal as professors and schools' managers. They can benefit if they have updated information about pupils number in each course, average grade for courses, schools' transport, lessons planning, absences planning, environmental data as atmospheric data, equipment of classrooms, teaching languages etc.

Next group of stakeholders are pupils and parents which are interested in transport conditions, weather conditions, available courses, ecological data as drinking water availability, radio and TV broadband and some SMART GIS infrastructure objects.

The third group are municipality and government employees in primary education sectors that use the data as number of pupils and students in each village and region, their classes, the teaching languages and their needs of equipment and professors. They have to minimize costs for transport, professors' engaging, costs for classrooms equipment and other material costs as heating costs, schools' maintenance at working conditions etc.

Transporters can have data for optimal routing, working time of schools, fuel supply, number of pupils and professors depending of locations etc.



Map

MAP MARKER

Geo Attribute

Latitude: Latitude

Longitude: Longitude

ADDITIONAL DATA

Color By: Row Count

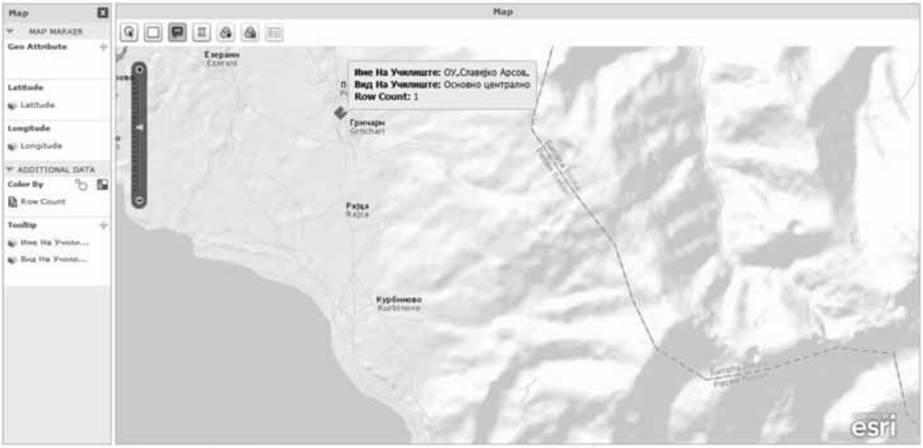
Tooltip: Име На Учили... / Вид На Учили...

Име На Училиште: ОУ„Димитар Влахов„

Вид На Училиште: Основно централно

Row Count: 1

Figure 4a – First layer of GIS-based VS4LEIS for the selected school in region



Map

MAP MARKER

Geo Attribute

Latitude: Latitude

Longitude: Longitude

ADDITIONAL DATA

Color By: Row Count

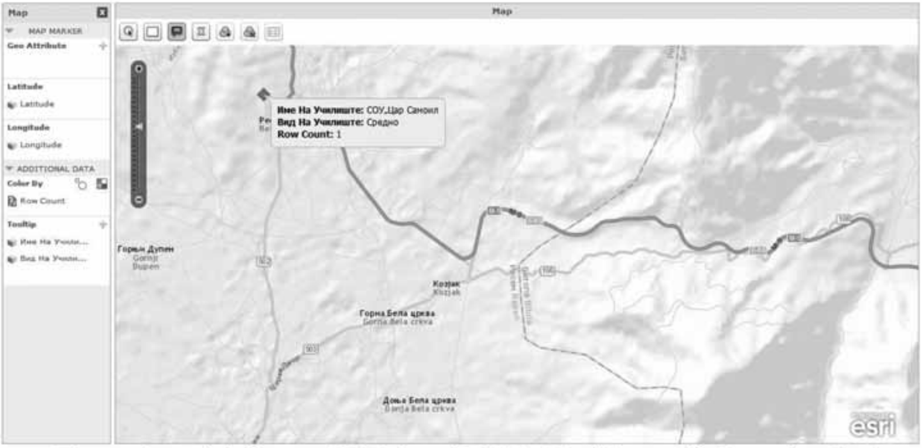
Tooltip: Име На Учили... / Вид На Учили...

Име На Училиште: ОУ„Славејко Арсов„

Вид На Училиште: Основно централно

Row Count: 1

Figure 4b – Second layer with presented other school' information



Име На Училиште: СОУ„Цар Самоил
Вид На Училиште: Средно
Row Count: 1

Figure 4c – Third layer of GIS-based VS4LEIS with travel information for the selected sub-region

6. Conclusions

Considering that the primary education is important pillar of education which introduces primary habits for using new technology trends among pupils at an early age and their professors, we start with modelling GIS-based VS4LEIS. Also, other stakeholders in primary education can use the proposed model for gaining benefits in decision-making processes as well as for gaining updated information. But, for implementation of this project, the trainings have to be provided for pupils and teachers in order to implement newest information technology and using of GIS-based VS4LEIS. This system can help to monitor data about the primary education level as well as all structures involved (government institutions, municipality institutions, schools, parents, pupils, students etc.).

One educational segment which is very important for gaining information and making habit of usage of the new technology is the primary education and for this reason this is the first point where IT has to be used in practical and efficient way. We modeled complex visual GIS-based system intended to stakeholders of dispersed schools through rural and other regions, taking into consideration many factors, data and legal obligations arisen from legal framework and transport costs as well as multilanguage support. The model is different from the central data model for primary education because of missing data caused from schools' dislocations, insufficient number of pupils, teaching staff and unavailability of local schools with missing road infrastructure etc. For this reasons, the model of GIS-based VS4LEIS is proposed, intended to local primary education stakeholders. We used business intelligence tool and visual data analysis intended to gain information for stakeholders in primary education.

The proposed model was created based on the gained prerequisites in survey made for this research. The survey was the base for the proposed model. Selection of the platform was made on the base of the stakeholders' demands and implementation processes was made on the selected rural region. Data collected from this research was imported in the selected software tools and some results are shown as separate layers. The next research activities is creating UML diagrams for the proposed model and for each group of stakeholders as well as creating other parts intended to each group of stakeholders with predefined rights and obligations for updating data and their usage.

The GIS-based VS4LEIS model development can be used as separate application in the school or as cloud service intended for usage from mobile or desktop devices and for this reasons the mobile services have to be specified and prepared as future research directions.

References

1. Aufmuth J., Centralized vs. Decentralized Systems: Academic Library Models for GIS and Remote Sensing Activities on Campus, Library Trends, Volume 55, Number 2, Fall 2006, pp. 340-347 | 10.1353/lib.2006.0000
2. Bednarz S.W. and Ludwig G., Ten things higher education needs to know about GIS in primary and secondary education,
3. Demirci A., Evaluating the Implementation and Effectiveness of GIS-Based Application in Secondary School Geography Lessons, American Journal of Applied Sciences 5 (3): 169-178, 2008, ISSN 1546-9239
4. Du Y, Yu C. Liu J. A Study of GIS Development Based on KML and Google Earth, Fifth International Joint Conference – Seoul, 25-27 Aug. 2009, Page(s): 1581 – 1585 E-ISBN : 978-0-7695-3769-6, Print ISBN: 978-1-4244-5209-5, DOI: 10.1109/NCM.2009.17, Publisher: IEEE
5. ESRI User Conference, 27.6.2016, http://video.esri.com/watch/5152/direction, accessed 20.7.2016
6. [4] ESRI 1995 ArcView White paper series: GIS in K-12 education, Redlands, Environmental systems researches research institute, inc.
7. Evandelidis K. and all, Geospatial services in the Cloud, Geocomputers & Gescience, Volume 63, February 2014, pg.116-122, doi:10.1016/j.cageo.2013.10.007
8. European Commission, 2007, European Commission Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), Off. J. Eur. Union, 50 (2007), pp. 1–14
9. EOSDIS, Earth Observing System Data and Information System, National Aeronautics and Space Administration. ⟨https://earthdata.nasa.gov/⟩ (accessed 23.07.2016).
10. Favier T.T. and van der Schee J.A., Exploring the characteristics of an optimal design for inquiry-based geography education with Geographic Information Systems, Computer and education, Volume 58, Issue 1, January 2012, Pages 666–677
11. Faggian A., McCann P., Human capital flows and regional knowledge assets: a simultaneous equation approach, Oxford Journals Social Sciences Oxford Economic Papers Volume 58, Issue 3, Pp. 475-500.
12. Forer P. & Unwin D., Enabling progression GIS and education, http://www.geos.ed.ac.uk/~gisteac/gis_book_abridged/files/ch54.pdf, 12.5.2016
13. Goetz S.J. and Rupasingha A., High-Tech Firm Clustering: Implications for Rural Areas, , American Journal of Agricultural Economics, Vol. 84, No. 5, Proceedings Issue (Dec., 2002), pp. 1229-1236, Published by: Oxford University Press on behalf of the Agricultural

& Applied Economics Association, Stable URL: http://www.jstor.org/stable/1245052, Page Count: 8

14. Gryl I., Jekel T., Re-centring Geoinformation in Secondary Education: Toward a Spatial Citizenship Approach, DOI: http://dx.doi.org/10.3138/carto.47.1.18
15. Johansson T. and Pellikka P., GISAS – Geographical Information systems applications for Schools. In: AGILE 2006, 9-th AGILE Conference on Geographic Information Science, Shaping the future on Geographic Information Science in Europe, pp.317-318, ALTO Press, Hungary,2006
16. Johansson T. , GISAS project: – Geographical Information systems applications for Schools, Finland: University of Helsinki (2006)
17. Johanson T., 2003, GIS in teaching education – facilitating GIS application in secondary school geography. ScanGIS'2003 On-line Paper, 285-293
18. Richardson D.B. and all., 2013, Spatial Turn in Health Research, Science, 22 March 2013, Vol.339, Issue 6126, pp. 1390-1392, DOI: 10.1126/science.1232257
19. Tokody D. & all., Smart City, Smart Infrastructure, Smart Railway, International Conference on Applied Internet and Information Technologies, 2015, pg.231-234
20. Wikle T.A., Finchum G.A., The emerging GIS degree landscape, Computer Environment and Urban Systems, Vol.27., Issue 2, March 2003, pg.107-122
21. Zhao P. and all, The Geoprocessing Web, Computers & Geosciences, Volume 47, October 2012, Pages 3–12, doi:10.1016/j.cageo.2012.04.021

Constructing Phylogenetic Networks and Trees for the Analysis of Genetic Datasets

Elitsa Kaloyanova[1]

[1]Saarland University, Saarbrücken, Germany

elitza_kaloyanova@yahoo.com

Abstract. The vast amount of data available due to new sequencing technology has presented an opportunity to explore processes from the field of chemistry, physics, biology and medicine in detail, but also raised the question of how they can be represented adequately. Representing the data as a set of rooted triplets and the construction of trees and networks from it has been proven useful in that aspect. This paper presents different data structures for constructing phylogenetic trees and networks, from a set of rooted triplets and will present rules regarding the relationship between the noise in the data and the complexity of the phylogenetic network.

Keywords: Bioinformatics, Phylogenetics, Rooted triplets

1. Introduction

Many scientific methods are dedicated to inferring a phylogenetic tree from a set of triplets [1,3]. However, constructing such a tree is not always possible and depends on the set of triplets. The challenge thereby lies in the fact that although many evolutionary relationships can indeed be represented accurately by a tree, there are some processes, such as horizontal gene transfer, modification of gene functions, chromosomal rearrangements, etc., where a tree structure does not suffice [17]. Moreover, there are some events where a tree representation is not adequate due to the level of complexity of the data. Furthermore, we need to account for the fact that the data might contain noise due to which a construction of a tree might not be possible. If this is the case, we can either build a tree with a subset of the triplets or choose a more complex structure to represent the entire set. The disadvantage in the former case is that some of the information, which may be valuable, is lost. In the later case, we choose to retain all the information, but represent it as a structure different than a tree. In this paper we explore the use of different tree structures to analyse phylogenetic data, as well as different algorithms for their construction. Our goal is determining the best possible representation of a rooted triplets set, based on the complexity of the data, by incorporating measures to decrease its level of noise and implementing an efficient data structure that accurately depicts the genetic relationships in the phylogenetic dataset.

The next part of the paper is organized as follows. In section 2 we introduce some preliminary definitions, such as the definitions of phylogenetics and rooted triplets, as well some of the related work in the construction of phylogenetic networks and trees. In section 3 we discuss the methods we use to construct networks and trees from sets of rooted triplets and tools for the analysis of the biological data. In the section 4 some results are presented and discussed.

2. Background

In order to construct a data structure, to represent a data set of rooted triplets, we first need to define the biological and algorithmic preliminaries pertaining to the construction of phylogenetic trees and networks. In this chapter we introduce some basics from the field of phylogenetics and graph theory. Furthermore we discuss previous work accomplished in this area.

2.1 Phylogeny

A Phylogeny describes the evolutionary history of a genetic set of taxa. [14] Phylogenetics is the study of phylogeny. Its main concern is to discover relationships between different species and find similarities between analog/related genes in different/same species [11,15]. It is of interest to find common ancestors of different organism and thus infer a timeline for the evolution of physical traits, gene or genome variation. One can always investigate the relationship between genes themselves, thus obtaining information about divergence, conservation or mutation rate of genes [17]. These give insight into the role of genes, how important or essential they are in the organism and help better understand the processes, undergoing in a living organism. One of the major branches of evolutionary biology is concerned with the evolution of viruses and their vast mutation rate. Due to their diversity and fast evolution, viruses are able to infect various hosts, jump across species, and it is difficult to find remedies against them [16]. Understanding the rates of virological evolution can give insight into how one can prevent infections and diseases from a viral source.

In the past years, due to the increase of the taxa available, it has become common praxis to analyze the data via construction of phylogenetic networks.

2.2. Algorithmic preliminaries

In this subsection we introduce definition from the field of graph theory, including the definition of a rooted triplet and a phylogenetic network.

Given any tree *T* , let *L(T)* be the leaf set of *T*. If *T* is a set of trees, let *L(T)* be the union of the leaf sets of the trees in *T*.
Let *r(T)* denote the set of rooted triplets that are induced subtrees of a rooted tree *T*, then the set *r(T)* is called the rooted triple set.

Def 1 Rooted triplet: A rooted triplet xy|z is consistent with a phylogenetic network, if ∃ a path from x to y which does not intersect the path from z to the root (note that here it is not required for this path to be unique) as is seen in Figure 1.

In order to describe the algorithm for constructing a level-1 phylogenetic network according to the algorithm in [2] we need to define a SN set. The name SN comes from subnetwork and, as each SN-set is a subnetwork in the final output of the algorithm and is defined as follows: For any X ⊂ L the set SN(X) is defined recursively as SN(X∪{c}) if there exists some x1,x2 ∈ X and c ∈ L\X such that ({x1,c},x2) ∈ T, and as X otherwise.

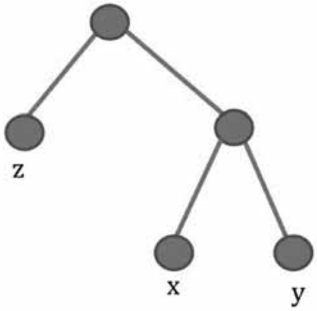


Figure 1. A rooted triplet

Def 2 Phylogenetic network: A phylogenetic network is a connected, rooted, simple, directed acyclic graph in which:

- each node has outdegree at most 2;
- each node has indegree 1 or 2, except the root node which has indegree 0;
- no node has both indegree 1 and outdegree 1;
- all nodes with outdegree 0 are labeled in such a way that no two nodes are assigned the same label.

Def 3 Span: Let R be a set of rooted triplets, then the span of R, or < R > is the set of rooted trees that are compatible with R and have leaf sets *L*(R).

Def 4 Closure: The closure of a consistent set R is:

$$\overline{R} = \bigcap_{T \in r<T>} r(T)$$

Def 5 NP: A language $L \subseteq \{0,1\}$ is **NP-complete** if

1. $L \in NP$, and
2. $L' \leq_p L for\ every\ L' \in NP$

If a language L satisfies property 2, but not necessarily property 1, we say that L is **NP-hard** [8].

2.3. Related work

There are many methods dedicated to constructing phylogenetic networks from rooted triplets [2,7,9,10,12,13]. In this subsection we discuss some of the related work and algorithms used to obtain phylogenetic trees and networks from data sets of rooted triplets.

The BUILD

Developed by Aho et al in 1981 [6], BUILD is a polynomial algorithm for the construction of a phylogenetic tree from a set of rooted triplets. The output of the algorithm is a phylogenetic tree, if one exists and NULL otherwise. The algorithm is based on the construction of the Aho et al [6] graph, also called auxilary graph and computing the connected components of the graph. The algorithm recursively computes connected components and if at any step there exists only 1 such component, the algorithm terminates with output NULL as the set of rooted triplets is inconsistent with any phylogenetic tree. Nevertheless, if no tree exists, it does not necessarily mean that no information is contained in the set.

Furthermore, the algorithm gives no indication as to what the level of inconsistency in the set is.

Algorithm FindClosure

Finding the closure cl(T) of a consistent set of rooted triplets T is an important task, as it displays all the information contained in T. FindClosure is a polynomial algorithm for determining the closure of a set of rooted triples with a complexity in O(m5), where m = |T| [2]. In the next chapter, we will propose an improvement on the algorithm, which results in better performance.

Inferring a level-1 Phylogenetic Network from a dense set of rooted triplets.

This method is only applicable in the case we can ensure that T is consistent, i.e. the BUILD algorithm returns a tree. However, we can not always ensure consistency and thus how one can infer a tree or the closure for an inconsistent set of triplets is also of interest. There exists a polynomial algorithm for constructing a level-1 phylogenetic network from a dense set of rooted triplets, as described by Jesper

Jannson [2].

3. Methods

Closure of a Level-1 network

So far we have discussed finding the closure of a set of rooted triplets only in the context of phylogenetic trees. We have seen that if we apply certain restrictions to the triple set, we can obtain the closure in polynomial time [5]. In this section we extend the definition of ***closure***(Def 3) to include phylogenetic networks and explore specific types of networks, for which computation of the closure can be achieved in polynomial time. Although having proven very popular in representing evolutionary data, phylogenetic trees are not always sufficient to accurately or fully represent the biological data. In particular, in cases of data containing noise or representing complex evolutionary events, such as hybridization, orthogonal gene transfer etc., a phylogenetic tree does not display the entire data set. Many algorithms have been devoted to finding an optimal subset of the input, in order to construct a phylogenetic tree. However, in many instances we cannot correctly identify the intercept in the data. Furthermore, if the data depicts a complex evolutionary process, a phylogenetic tree does not provide an accurate representation of it. Therefore, a phylogenetic network has proven to be a more suitable mean for representing phylogenetic data of the above mention texture.

The problem here is that finding the complexity (level) of the network to be built from a set of rooted triplets is NP-hard (Def 4).

Depending on the connectivity of the auxiliary graph it is possible to determine whether a tree can be built or not. However this is a measure only for the basic tree case (level-0-network) and does not reflect the connection between noise and a higher level network.

Due to the complexity of the problem and the running time increase associated with it, it is impractical to construct a network of high complexity. Therefore our goal is to implement a network with low k, by introducing noise reducing measures in the preprocessing stage of the analysis.

We implement an algorithm for the construction of a level-1 phylogenetic network, when the data set, contains a closed set of rooted triplets. In each step the algorithm computes the connectivity of the auxilary graph, by consecutively adding rooted triplets. If no inconsistencies are found the output of the algorithm is a phylogenetic network, where the first and last end nodes represent the data with the furthest connection in the set, and adjacent nodes exhibit the closest relationship in the dataset. When the data set is not closed, we implement different methods, to reduce the noise and subsequently the set to a closed one.

The first method searches the data set for sets of 2, 3 and 4 conflicting triplets, removing one of the triplets in order to obtain a closed set. The threshold for allowed removals can be user defined, but based on observation from our tests and experimental evidence, the majority of data sets contain between 5 and 15 percent noise, therefore, our threshold is set at 15 percent.

The second approach includes a scoring system, based on weighted nodes in the triplets, whereby triplets, containing a higher score, achieve more confidence, compared to a lower scored triplet. In this case when a conflict emerges, the lower scoring triplet is removed from the set. This method is effective, when the user has prior knowledge of the data set or is interested in particular genes or species.

4. Experiments and Discussion

In order to test our algorithms, we applied them on different data sets of rooted triplets, varying in size and in complexity, concerning the number of different genes and the level of noise contained in the sets.

The test data consists of randomly generated datasets, containing 50 to 100 sets of triplets. Furthermore we tested the algorithm on one set of 87 enzymes from the cytochrome family, obtained from the uniprot library, as well as another set of expression rna data, from the ncbi database.

When we applied the first procedure for accessing the noise in the data set, we constructed a level-1 network or a level-0 network on average with 40% more of the sets.

In the cases, where we had prior information for the genes or obtained a weighted triplet set, we were able to use the second measure and obtained a level-1 network in 35% more of the cases. Our observations are that in this case, the probability of removing a false negative from the set decreases compared to the first method of elimination, due to the additional information provided from the experimental data. However, in many cases we do not obtain any further information on the genes, in which case we suggest the use of the first measure.

When using datasets of similar genes of different species we observed many homologue genes, coding for genes and expression of rna and proteins and enzymes, connected to cancer, members of the cytochrome family, enzymes connected to the oxidation of organic substances. There is a strong similarity between many of the genes in a mouse and human dna and construction of a phylogenetic network allows for an easy and fast comparison and detection of orthologue genes, which can be useful in developing a drug therapy and deciding on the phase 2 model organism.

The algorithm was performed on a 64 bit unix system with a Intel Xenon processor W3540 with 2.93Ghz and 12GB of RAM. All other programmes were performed on a 32 bit unix system equipped with an Intel Atom processor with 1.66Ghz and 2GB of RAM. For smaller and medium size sets the average time for constructing the level-1 network is in O(m2), for larger data sets it is in O(m3).

Conclusion

In this paper we introduce different data structures for the representation of phylogenetic data, in particular, sets of rooted phylogenetic triplets. Moreover we introduce measures for the assessment of the level of noise, such a data set may contain and propose stochastic methods for its decrease in the preprocessing stage of the analysis.

We observe that incorporating a preprocessing step in the analysis of the data can increase the accuracy of the method for constructing a phylogenetic network. However, the increase in size of the dataset also affects the complexity of the data structure, required to incorporate it. This is due not only to the amount of data contained, but also to the elevation of noise and inconsistencies in the set, which correlate with the dimensions of the data.

As future work we would like to find an efficient algorithm for constructing a level-k phylogenetic network for a fixed k.However, as the k increases, so will the complexity of the algorithm and therefore for k¿3 no efficient algorithm exists for creating a level-k network.

Another improvement would be to find a way to decide on the complexity of the dataset before running the algorithm. If we could assess in a preliminary step whether a data set would result in a level-1 or level-5 network we could use much more efficient algorithms. So far we have found no concrete indicators as to how complex the network representation of the set actually is.

References

[1] Jaroslaw Byrka, Sylvain Guillemot, and Jesper Jansson. New results on optimizing rooted triplets consistency. Discrete Applied Mathematics, 158(11):1136–1147, (2010).

[2] Jesper Jansson and Wing-Kin Sung. Inferring a level-1 phylogenetic network from a dense set of rooted triplets. Theor. Comput. Sci., 363(1):60–68, (2006).

[3] J. Felsenstein. Inferring Phylogenies, Sinauer Associates, Inc., Sunderland, (2004).

[4]. Huson DH, Rupp R, Scornavacca C Phylogenetic Networks Concepts, Algorithms and Applications. Cambridge University Press, (2010) .

[5] J. Jansson, A. Lingas, and E.-M. Lundell. The approximability of maximum rooted triplets consistency with fan triplets and forbidden triplets, Proceedings of CPM 2015, Lecture Notes in Computer Science, Vol. 9133, pp. 272–283, (2015).

[6] A. V. Aho, Y. Sagiv, T. G. Szymanski, and J. D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions, SIAM Journal on Computing, Vol. 10, No. 3, pp. 405–421, (1981).

[7].To TH, Habib M Level-k Phylogenetic Networks are Constructable from a Dense Triplet Set in Polynomial Time. In CPM09 5577: 275–288. doi: 10.1007/978-3-642-02441-2_25, (2009).

[8]. Cormen, T.T.; Leiserson, C.E.; Rivest, R.L. Introduction to algorithms; MIT Press: Cambridge, 249 MA, USA, (1990).

[9]. Wu, B.Y. Constructing the Maximum Consensus Tree from Rooted Triples. Journal of 232 Combinatorial Optimization, 8, 29–39, (2004).

[10] Shuying Li Phylogenetic Tree Construction using Markov Chain Monte Carlo Fred Hutchinson Cancer Research Center August, (1996).

[11] C. Jill Harrison and Jane A. Langdale A step by step guide to phylogeny reconstruction The Plant Journal 45, 561–572, (2006).

[12] Ronnie Bathoorn, Arno Siebes. Constructing (Almost) Phylogenetic Trees

from Developmental Sequences Data. European Conference on Principles of Data Mining and Knowledge Discovery, (2009).

[13] Jie Yang 1 , Zhi Cao , Huanwen Chen , Kai Long , Gangcheng Li , Li Zhao A Method for Constructing Phylogenetic Tree Based on the Minimum Spanning Tree of the Complete Graph. MATCH Commun. Math. Comput. Chem. 65 469-476, (2011).

[14] Bruce Alberts, Alexander Johnson, Julian Lewis,Martin Raff, Keith Roberts, Peter Walter Molecular Biology of the Cell Fifth Edition Garland Science, (2008)

[15] William S, Klug, Miachel R. Cummings, Charlotte A. Spencer Genetik 8., aktualisierte Auflage Pearson, (2007).

[16] Nicholas H. Acheson Fundamentals of Molecular Virology. 2nd Edition Wiley, (2011).

[17] Brenda A. Wilson, Abigail A. Salyers. Dixie D. Whitt, Malcolm E. Winkler, Bacterial Pathogenesis A Molecular Approach, Third Edition ASM Press, (2011).

Gamification Systems to Support the Learning Process

Teo Petrov[1], Kalinka Kaloyanova[1, 2]

[1]Faculty of Mathematics and Informatics, Sofia University, 5 James Bourchier blvd., 1164, Sofia, Bulgaria
[2]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. Georgi Bonchev Str., Block 8, 1113, Sofia, Bulgaria

teopetrov@gmail.com, kkaloyanova@fmi.uni-sofia.bg

Abstract. Implementing methods and techniques of gamification can support the learning process of students in a university by increasing their interest and enhance the impact of their studies. Using such practices deepens understanding of the material and the final result is better quality in the teaching process. In this paper a comprehensive analysis of the subject area has been accomplished. Based on the research, main requirements and functionality of a system implementing methods and techniques of gamification are defined.

Keywords: Information Systems, Gamification, Gamification in Education

1. Introduction

The education of students is an extremely important process which requires enhancements in all possible ways. A number of practices and methods are proven to operate at different levels, even on business one [13], but they can be supported with new ideas, following the latest trends in information technologies.

In this paper we explore requirements and functionality of a system that uses the methods of gamification to support the process of student's education in Information systems (IS) undergraduate program in Sofia University - Faculty of Mathematics and Informatics (FMI). To achieve this, we first carry out a comprehensive and detailed analysis of the subject area, considering gamification as a tool to improve the results of the teaching and how the effects of the educational process and the interest of students can be enhanced. Based on this study the basic functionality of such a system is proposed.

2. Gamification

The term "gamification" is entering more and more common in the vocabulary of companies. The main reason for integrating this approach is the building of

V. Dimitrov, V. Georgiev (Editors): ISGT'2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

strong loyalty of users, be they customers or employees. The key to success is that it is derived based on the high level of activity and participation that build emotional commitment and long-term relationship “user-system” that is much stronger and powerful than standard approaches.

Gamification represents the concept of using gaming techniques and mechanisms in various areas and work processes. The purpose of gamification is not to create a game, but to offer a sense of game elements to the consumers which will inspire and motivate them [9]. To achieve this, the basic ideas of games like entertainment, goals, awards, challenges, hierarchy, rankings and similar are taken and applied to real goals instead of just for entertainment [1].

The types of rewards are usually virtual points, badges or achievements that may eventually be exchanged for awards in real life. Making achieved by a user visible to others and maintaining the rankings between players, creates a strong competitive feeling between users and increases the desire for competition among them. Since these incentives have long been used and are already fully proven in the gaming industry it would be only useful if they are used in other areas with the same success.

2.1 “Activity-Achievement-Sharing”

The “Activity-Achievement-Sharing” loop (Fig.1) is a key point for gamification. Its purpose is to use the internal motives of users which are natural to humans and therefore work much more efficiently than external ones such as assigned rules, requirements and objectives.

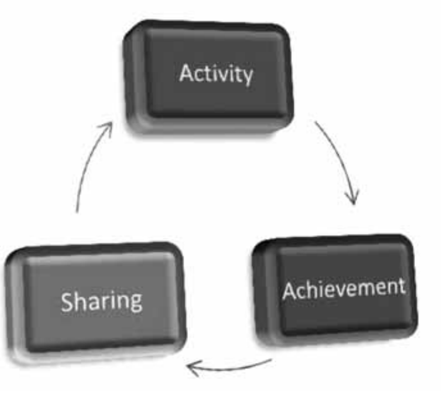


Fig.1. The “activity-achievement-sharing” loop.

The activity of users can be demonstrated using a performance indicator of assigned tasks, diversity in the selection of possible assignments, speedy feedback on their work and other activities that come within their competence.

Mainly, the achievements are based on internal satisfaction of people. Material rewards are not prohibited, but they are not the fundamental idea in gamification. The apparent progress of human activity and obtainable awards are strong internal stimulators.

2.2 Main principles for successful gamification

In order to achieve the main objectives of gamification - higher engagement of users, encouraging innovation and directing the user behavior in the desired direction, a few basic principles could be distinguished [2]:

- **Accelerated feedback loop**

  In standard applications, verification of the result of user activity is too slow. The distances between major check-ups are periods during which the user is likely to give up or at least lose interest.

- **Clear "game" concept and clear rules**

  Users should be encouraged to pursue individual steps of the game, which lead to the big goal. Vague, confused and complicated rules would confuse them and reduce their interest. In the opposite case, clear objectives and rules motivate users to continue in the "game".

- **Defiant description**

  The creativity of the team responsible for gamification must occur at this point because one of the main ideas in the games is the temptation and challenge of exciting tasks and the final result. Only appropriate presentation will encourage users to participate and fulfill the tasks set by the architects of gamification.

- **Appropriate tasks**

  The complexity must be selected very precisely. It should not be too complicated because it will despair most users and they will give up. Also the tasks cannot be too easy because you will not awake the desire to race in people with something boring. A good approach is to divide the big goal into separate smaller tasks, that are appropriately selected and distributed.

2.3 Using Gamification in different areas

A Forbes study [4] predicted that by 2014 at about 70% of companies will implement gamification systems in order to improve their marketing and customer retention. Nowadays, gamification is even more widespread and its market is expected to raise to near $2.8 billion by the end of 2016 [5].

A good example of a possible application of gamification would be in academic activities of a university [6].

Using gamification can improve the educational activities. One of the most important advantages, which gamification gives to students, is the feeling that they control the learning process. Another important aspect is the creation of an objective to pursue. Thus training turns from attending lectures and taking an exam, into a long process of fulfilling criteria of knowledge, completing individual tasks and participating in discussions of problems. This is a way to unnoticeably get insight into the essence of the subject matter.

The introduction of ranking between students, based on earned points, can play a very motivating role. As the learning process anyway implies evaluation, other approaches, like described in [10] use the opportunity to make rankings of students in real time using such systems. But the implementation of gamification is convenient time to enhance these processes.

2.4 Methods and techniques

The successful integration of gamification and achieving a positive effect require the use of a large set of methods and techniques. Over the years a large number of resources have been evolved. Some of the methods and techniques are psychological, while others focus more on the technical part, so in each individual case the architects of the gamification must determine the optimal approach, interweaving different methods and techniques.

Based on [3] some of main methods, mechanisms and techniques, which could be used in the educational area, are listed below:

- Achievements - these are virtual or real evidence that the user has completed a certain task. The tasks can be of various nature and difficulty, as well as assigned to team or individuals. The achievements are a way for the “players” to demonstrate to the other users what they have achieved. Some of the possible achievements initially are “locked” and cannot be seen before reaching a specific set of tasks, which creates additional excitement and stimulates the user.
- Appointment dynamics - the system expects the user to perform a certain task at a specific date and time.
- “Blissful” Productivity - to ensure that the users feel that they are doing something productive and useful, which will bring them satisfaction afterwards. The pursuit to feel happiness playing is a basic human aspiration, which is proven by the huge success of the games.
- Bonuses - awards following the completion of a series of challenges or other main tasks. They could be given for completion of the “combo”, any specific task or another very appropriate option - a quality performance.
- Combo - a specific term from the gaming world, describing the idea of performing a combination of tasks that leads to an additional award.

Besides that it adds enthusiasm to work, the combos create a suitable atmosphere that encourages the continuation of the work. A combo usually leads to the receiving of a bonus.

- Levels - separate steps on which the user “climbs” covering certain criteria, most often it comes to accumulating virtual points. Often, certain features of the system are unlocked when the users reach a certain level, which encourages the wish to gain more virtual points. The levels are proven to be one of the strongest motivators for players and apply to almost any version of gamification.
- Loss aversion - opposite of the standard approach to stimulate progress in gamification option is the creation of light penalties. Their purpose is more to hold the user’s attention on the system than actually to reduce his results. The possibility of losing status, ranking or opportunity, for which the player has fought for months, will keep them “playing”.
- Points - one of the techniques that is almost mandatory for gamification. They are a universal tool for measuring the results of the players’ progress achievement and determine the ranking table of users. The points represent a virtual value, which describes the progress of the user in the system. For every achievement or bonus, the user receives a certain number of points. Usually they are the only way to move to the next level. They also have great value for architects and developers of the system as an objective indicator of the progress of individual users and can be used in various cases.
- Progress - dynamic representation of the progress of the user and his achievements and points. It is usually closely related to the levels and the points because in practice it depends on them.
- Status - a specific way to guarantee the effort of the user to progress is the use of individual statuses. Points and badges are used to achieve different statuses that show the strengths and achievements of a given user.

3. Using Gamification in Education

The gamification can be integrated in academic activities with the primary goal to improve the quality of teaching students as well as to increase their interest in learning new material.

In [12] we described a specific implementation of the project approach in students education for undergraduate IS program at FMI - Sofia University. Now we focus on using game elements to evaluate the individual student’s contribution in team projects at different courses.

This goal could be achieved by relying on the main gamification method -

involvement in a game situations, which includes completing tasks, winning and sharing awards, stimulated feedback between students and teachers through comments, participation in forum and others gamification techniques. A variety of connections and options for sharing achievements could be used to seek natural stimuli such as a desire for victory, knowledge, experience and skills obtained, as well as satisfaction from a job well done.

It is also useful to see some statistics of students involvement in the various activities of the courses. These data could be used to produce valuable analyzes for the lecturers to make any necessary corrections. It would also be of great help in the creation of next courses to enable teachers to comply with perceived trends and preferences among students.

Below the main elements for an university information system with gamification elements are discussed.

3.1 Users

Any information system that relies on gamification to support the education, should support several important roles - two types of teachers (lecturers and assistants), students (the most common and standard user),and a system administrator.

Lecturers initially create courses, provide and distribute assignments and prepare the overall plan for the course.

Assistants have the same responsibilities as at the learning process in real life - audit and evaluation of students' homeworks and other tasks. The assistants are appointed by the lecturer of the course.

Students are the most common users of the system. Each student who attended the course has the corresponding user in the system. The students upload their homeworks and other assignments and check their results. They should have detailed access to the competitive functions of the system, as well as information on the current tasks.

The role of the *system administrator* as usual is to create and manage the accounts of all users and to process written requests for use of the system.

3.2 Functionality

The basic functionality of a such system could spreadin several directions.

- Course management

The maintenance of the courses is supported by lecturers. A lecturer can create different courses. When creating a course, he should identify the other lecturers/assistants of the same course.

- Students enrollment

All students should have the ability to check the possible courses, view their descriptions, and discuss information about them in the comments. If a student decides to attend a course, he reflects in the system to be able to participate fully in and use all advantages that the current application gives. A student not only has the right but he is expected to participate in many courses.

- Assignments

The system should support different assignments (homeworks, tasks, etc.) with appropriate rewards as received points. Therefore, each of the assignments has a maximum number of points that can be obtained. Thus the evaluation process is a much fairer and more natural way to prevent situations in which a student always chooses the easiest assignments and works least. The submitting of every homework is done only by the student himself. The evaluation of homework can be done by the lecturers or assistants.

- Feedback

The feedback is the ability of system users to express their opinions and participate in discussions on various topics related to their activities. For this purpose, a forum, in which they can create separate topics, to write posts , and express their opinion should be available. There are no technical or other restrictions on who can use these functions of the application and practically every registered user has these rights. Some comments on different parts of the information system (courses, homework, badges and awards) could also be made.

- Administrative management

The administrative management mainly concerns the administration of user profiles. As usual this functionality is covered by the system administrator.

3.3Achievements

The most specific part of the application is to support the achievements that students can earn by their active participation in the course - badges, awards and ranking tables. Participating in the course, students periodically perform tasks, whose rates became major part of their final grade. Achievements are provided just for these tasks according to the course specificity. When students meet certain criteria to receive a badge or award, they don't own it forever and could lose it if their performance falls below a certain level.

The management of achievements is usually performed by the lecturer of the course.

Achievements are divided into two main categories:

- Awards - actual prizes, for which students are contending. Obtained for significant achievements in assignments and activities of the course. Winning awards, the students receive bonus points, which effectively represent an increase of their grade in a particular range.
- Badges - these are secondary achievements that are obtained for less difficult tasks. They are used to encourage the participants.

The badges and awards usually are handed out automatically by the system and can be seen in the user's profile.

The scoring table is the other main method of gamification. It provides various options for review and analysis of the achievements of the students. Basically, it provides ranking among the students built based on points earned by homework and achievements, but also allows to filtrate data.

The scoring table is formed using every student's points, which are combination of homework marks and achievements. This way it becomes a reflection of the success of students in the course, as points can be earned through quality homeworks and winning awards.

Table ranking is also essential for the course as it plays role of a powerful natural stimulant and creates great sense of competition between users. The table is visible to all users to encourage stronger ambitions.

3.4 Main requirements

A Gamification educational system should provide basic functionality needed for successfully maintaining university courses in which students are actively involved in the academic activities. Because such a system is too complex for designing and developing, it is not easy to determine all requirements at once, so we decided to use theMoSCoW method [7] to formulate them. In this way the system also could be developed incrementally.

Following the MoSCoW principles we divided the main requirements of a such information system by their significance in four categories:

- Basic services that the system **must have**:
    - Course management- creating, updating deleting courses
    - Homework management - creating, uploading and rating homeworks
    - Achievements management (both badges and awards)
    - Scoring table – based on earned points of every student. Students can enroll and leave courses (or forced to leave by university administrators).
    - Account management.
- In addition the system **should have**:
    - Homeworks – comment section for users and restrict some file types.
    - Exam Module
    - Notification support
    - Sharing in social networks – integration with popular social networks and potential university network
- Also the system **could have**:

    - Forum
    - Activity log
    - Messaging between users
    - System for bonus points for active students
    - Detection of copied texts for homework
- At this level **won’t have**:
    - Connection with social networks’ mobile applications
    - User registration with social network profile
    - Chat
    - System for mutual support between students
    - Newsletter
    - Poll system.

4. Conclusion

In this paper we discussed how the implementation of methods and techniques from the field of gamification into a specific information system can support the learning process of students in a university by increasing their interest and strengthen the impact of their studies.

Various concepts in gamification area have been discussed and described to focus on the main characteristics of gamification.

Based on the research, basic functional requirements for an information system which follows the idea of gamification to support the learning process are defined. Our next steps will be to develop such a system to support the education in the IS undergraduate programme at FMI - Sofia University, following the above consideration as well some of the recommendations discussed in [8], [9] and [11].

Acknowledgment. This paper is supported by Sofia University “St. Kliment Ohridski” SRF under Contract 68/2016.

References

1. Karen Robson et al: Is It All a Game? Understanding the Principles of Gamification, Business Horizons, vol. 58, Issue 4, pp 411-420 (2015)
2. Oprescu F., Jones C., Katsikitis M.: I Playat Work—ten principles for transforming work processes through gamification, Frontiers in psychology, 5, doi.org/10.3389/fpsyg.2014.00014 (2014)
3. Game mechanics, https://badgeville.com/wiki/Game_Mechanics
4. L. Goasduff: Gartner Enterprise Architecture Summit, London (2011)
5. “Gamified Engagement”, http://m2research.com/Gamification.htm
6. Jackson G. T., McNamara D. S.: Motivation and performance in a game-based intelligent tutoring system, Journal of Educational Psychology (2013)

7. Miranda E.: Time Boxing Planning: Buffered Moscow Rules, http://mse.isri.cmu.edu/software-engineering/documents/faculty-publications/miranda/mirandabufferedmoscowrules.pdf
8. Boytchev, P., Armyanov, P.: Re-experiencing Engineering Inventions within a Modern Virtual Environment, In Proceedings of 2nd International Conference Software, Services & Semantic Technologies S3T 2010, Varna, Bulgaria, pp. 55-62 (2010)
9. Dicheva D., Dichev C., Agre G., Angelova G.: Gamification in Education: A Systematic Mapping Study, Educational Technology & Society, 18 (3), pp. 75-88 (2015)
10. Manev Kr, Sredkov M., Armyanov P.: Software Platform for Teaching Programming With Grading Systems, In Proceedings of the Fortieth Jubilee Spring Conference of the Union of Bulgarian Mathematicians, Borovetz, pp. 300-305 (2011)
11. Kanabar V., Chitkushev L.: Innovative Applied Certificates in Computer Information Systems for Undergraduate Students , In Proceedings of the 9th Annual Conference "Computer Science and Education in Computer Science", Fulda-Würzburg, Germany, pp. 93-95 (2013)
12. Kaloyanova K., An Implementation of the Project Approach in Teaching Information Systems Courses, In Proceedings of the 8th International Technology, Education and Development Conference, INTED 2014, Valencia, Spain, pp. 7090-7096 (2014)
13. Shahinyan K., Krastev E., Evaluation Metrics for Business Processes in an Academic Environment, In Proceedings of the 7th Int. Conference Information Systems & GRID Technologies, Sofia, Bulgaria, pp. 297- 306 (2013)

COBBIS system cataloging – review

Blagojce Najdovski[1], Violeta Manevska[2], Snezana Savoska[2]
[1]Faculty of Biotechnical science, Bitolska bb,
University „St.Kliment Ohridski“ – Bitola,7000, R.of Macedonia,
[2]Faculty of information and communication technology, Bitolska bb,
University „St.Kliment Ohridski“ – Bitola,7000, R.of Macedonia,
bnajdovski@yahoo.com, violeta.manevsks@fikt.edu.mk, snezana.savoska@fikt.edu.mk

Abstract. In the today's information society libraries are resource centers of knowledge that also have purpose to run entries planned results obtained from some intelligent systems. In early 1987 was first adopted a common system for cataloging by in this time community of the Yugoslav National Libraries as a common ground for the library information system and scientific and technological information of Yugoslavia. The management and operation of the system is taken from the Institute of Computer Science in Maribor. In 1991, IZUM (Institute of Information Science Maribor), promoted COBISS as upgrade and replacement for cataloging. Due to the collapse of Yugoslavia, libraries outside Slovenia gave up from usage of this system and gave resignation from the library system and began to work independently based on the COBISS platform, a shared cataloging network COBISS.net COBISS.SI, COBISS.SR, COBISS.MK , COBISS.BH + COBISS.RS, COBISS.CG, COBISS.BG and COBISS.AL tags autonomous library systems are built in certain countries.

Keywords: COBBIS, National Libraries, Institute of Computer Science, Cataloging System

1. Introduction

COBISS can be presented as an organizational model of joining libraries in the national library system with mutual cataloging system including local bibliographic databases, databases etc.

Functionalities that the system enables the following:

- Standardized and shared cataloging of library materials
- Suitably qualified catalogs
- Linking of libraries via computer and communications network.

Each national library system based on the COBISS platform was created by the Central Office of COBISS (NCC).

What can COBISS (NCC) is the following:

- Plan and coordinate activities related to the linking of libraries in the system.
- Provision of computing facilities for the operation of the system and central services.

- Provide program support of COBISS and replacement instructions cataloging.
- Training and professional help to libraries and other users of the program support of COBISS.
- Professional assistance between libraries with data transmission.

2. COBBIS Platform

COBISS3 is the third generation of software for library automation and access to different databases developed by IZUM. Server-side, the servers as the operating system use Windows, and Java-based application servers. The client side using Java based application that allows various desktop and laptop computers on different operating systems to function. As for the Internet connection to be able COBISS3 work without any difficulties it is the minimum speed of the Internet to reach 256 / kbps.

COBISS3 software architecture consists of the following parts:

- Graphical User Interface
- Data entry and search
- Using the ISO (10646)
- Flexibility
- Allows multilingualism

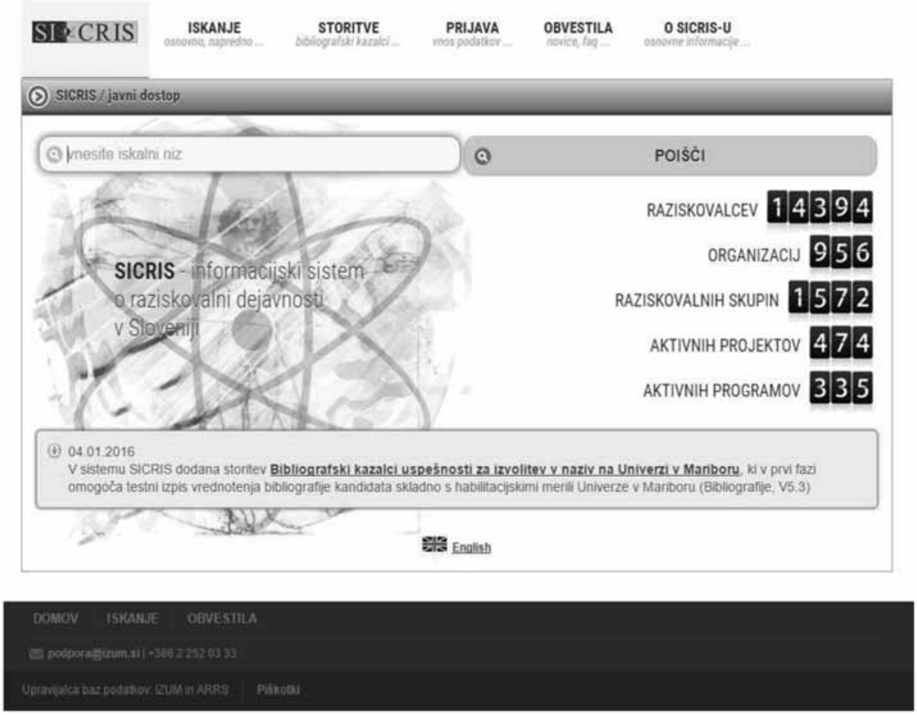
5. Mutual cataloging

Mutual cataloging enables a rational division of labor and time saving when managing catalogs. Each new item is cataloged only once, then all registered participants can access it. Below mutual cataloging means communication between the local database of individual libraries and common database cataloging. To exchange data in COBISS system there are several types of formats that are used as follows:

- COMARC / B - used for the exchange of bibliographic data.
- COMARC / A - used to exchange copyright data.
- COMARC / H - used to exchange certain specified data

4. Comparison of COBBIS systems

Repository of universities in Ljubljana COBISS.SL is connected to the system through the application of SICRIS. SICRIS system is developed and maintained by the Institute of Information Sciences in Maribor and Research Agency of the Republic of Slovenia. All information containing SICRIS system are integrated into a European research system information named ERGO connecting all European information systems together. The biggest advantage is in the search, where you can see various projects.



SICRIS - informacijski sistem o raziskovalni dejavnosti v Sloveniji

RAZISKOVALCEV 14394

ORGANIZACIJ 956

RAZISKOVALNIH SKUPIN 1572

AKTIVNIH PROJEKTOV 474

AKTIVNIH PROGRAMOV 335

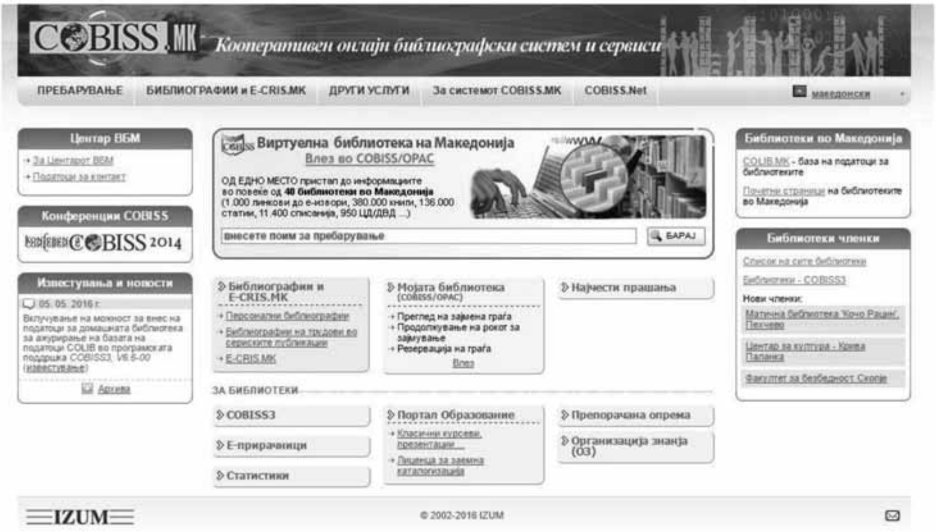
04.01.2016
V sistemu SICRIS dodana storitev **Bibliografski kazalci uspešnosti za izvolitev v naziv na Univerzi v Mariboru**, ki v prvi fazi omogoča testni izpis vrednotenja bibliografije kandidata skladno s habilitacijskimi merili Univerze v Mariboru (Bibliografije, V5.3)

podpora@izum.si | +386 2 252 03 33

Upravljalca baz podatkov: IZUM in ARRS

Figure 1. .COBBIS Maribor

COBISS system is used in R. Macedonia and its connecting all libraries in Macedonia together. The look and functionality of the COBISS system are simple. COBBIS system contains different search filters depending on the need for search. The search can be done by: author, title of the book, most popular, newest etc.. The system has the capabilities for a broader search. On the website of the system displays all libraries that are part of the system or that of its members, so that it can perform search by a specific library. The main search system in Macedonia is enabled through COLIB.MK where doing a search according to the required publication.



Вклучување на можност за внес на податоци за домашната библиотека за ажурирање на базата на податоци COLIB во програмската поддршка COBISS3, V6.6-00 (известување)

ОД ЕДНО МЕСТО пристап до информациите во повеќе од **40 библиотеки во Македонија** (1.000 линкови до е-извори, 380.000 книги, 136.000 статии, 11.400 списанија, 950 ЦД/ДВД ...)

Figure 2. .COBBIS Macedonia

4.1. Local applications

When it comes to local application refers to when he entered a certain amount of data and presentation of data to end users. The data is processed and displayed in local applications such COMARC / H. Popular local applications used are the following: COBISS3 / Acquisitions, COBISS3 / Serials, COBISS3 / Holdings, COBISS3 / Loan, COBISS3 / Interlibrary, COBISS3 / Reports, COBISS3 / Application Administration.

5. Conclusions

With COBISS system facilitated the work in search of certain publications without having to spend time. Advantages offered COBISS system are great, though it is possible to easily search enabled and protection of all publications published by their illegal copying. A big advantage of this CMS is that can connecting libraries from different countries in Europe that have a subscription and whit this can easy access to information and books from any library that is part of the COBISS system.

References

1. http://www.cobiss.si/

2. http://home.izum.si/cobiss/oz/HTML/OZ_2015_2_final/index.html#6

3. http://www.vbm.mk/cobiss/

Detailed Formal Specification of Software Weakness CWE-128 (Analyzes)

Vladimir Dimitrov,

Faculty of Mathematics and Informatics, University of Sofia, 5 James Bourchier Blvd.,
1164 Sofia, Bulgaria
cht@fmi.uni-sofia.bg

Abstract. Software weaknesses are described in formatted text. There is no widely accepted formal notation for that purpose. This paper shows how Z-notation can be used for formal specification of CWE-128. The weakness is specified at different detail levels and informal level environment are discussed.

Keywords: software weakness, formalization, Z-notation.

1 CWE-128: Wrap-around Error

CWE-128 full description taken from [1] is given in the subsections below.

Description

Description Summary: Wrap around errors occur whenever a value is incremented past the maximum value for its type and therefore "wraps around" to a very small, negative, or undefined value.

Time of Introduction

- Implementation

Applicable Platforms

Languages:

- C: (Often)
- C++: (Often)

Common Consequences

Scope	Impact
Availability	Technical Impact: *DoS*: crash / exit / restart; *DoS*: resource consumption (CPU); *DoS*: resource consumption (memory); *DoS*: instability This weakness will generally lead to undefined behavior and therefore crashes. In the case of overflows involving loop index variables, the likelihood of infinite loops is also high.

Integrity	Technical Impact: Modify memory If the value in question is important to data (as opposed to flow), simple data corruption has occurred. Also, if the wrap around results in other conditions such as buffer overflows, further memory corruption may occur.
Confidentiality Availability Access Control	Technical Impact: Execute unauthorized code or commands; Bypass protection mechanism This weakness can sometimes trigger buffer overflows which can be used to execute arbitrary code. This is usually outside the scope of a program's implicit security policy.

Likelihood of Exploit

Medium

Demonstrative Examples

Example 1

The following image processing code allocates a table for images.

Example Language: C (Bad Code)

```
img_t table_ptr;
/*struct containing img data, 10kB each*/
int num_imgs;
...
num_imgs = get_num_imgs();
table_ptr = (img_t*)malloc(sizeof(img_t)*num_imgs);
...
```

This code intends to allocate a table of size num_imgs, however as num_imgs grows large, the calculation determining the size of the list will eventually overflow (CWE-190). This will result in a very small list to be allocated instead. If the subsequent code operates on the list as if it were num_imgs long, it may result in many types of out-of-bounds problems (CWE-119).

Potential Mitigations

Requirements specification: The choice could be made to use a language that is not susceptible to these issues.

Phase: Architecture and Design

Provide clear upper and lower bounds on the scale of any protocols designed.

Phase: Implementation

Place sanity checks on all incremented variables to ensure that they remain within reasonable bounds.

Background Details

Due to how addition is performed by computers, if a primitive is incremented past the maximum value possible for its storage space, the system will not recognize this, and therefore increment each bit as if it still had extra space. Because of how negative numbers are represented in binary, primitives interpreted as signed may "wrap" to very large negative values.

Weakness Ordinalities

Ordinality	**Description**
Primary	(where the weakness exists independent of other weaknesses)

Relationships

Nature	**Type**	**ID**	**Name**	**V**
ChildOf	C	189	Numeric Errors	699
ChildOf	Ⓒ	682	Incorrect Calculation	699 1000
ChildOf	C	742	CERT C Secure Coding Section 08 - Memory Management (MEM)	734
ChildOf	C	876	CERT C++ Secure Coding Section 08 - Memory Management (MEM)	868
ChildOf	C	998	SFP Secondary Cluster: Glitch in Computation	888
CanPrecede	Ⓒ	119	Improper Restriction of Operations within the Bounds of a Memory Buffer	1000
PeerOf	Ⓑ	190	Integer Overflow or Wraparound	1000

Relationship Notes

The relationship between overflow and wrap-around needs to be examined more closely, since several entries (including CWE-190) are closely related.

Causal Nature

Explicit

Taxonomy Mappings

Mapped Taxonomy Name	**Node ID**	**Fit**	**Mapped Node Name**
CLASP			Wrap-around error
CERT C Secure Coding	MEM07-C		Ensure that the arguments to calloc(), when multiplied, can be represented as a size_t
CERT C++ Secure Coding	MEM07-CPP		Ensure that the arguments to calloc(), when multiplied, can be represented as a size_t
Software Fault Patterns	SFP1		Glitch in computation

Related Attack Patterns

CAPEC-ID	**Attack Pattern Name** (CAPEC Version: 2.8)
CAPEC-92	Forced Integer Overflow

References

[REF-17] Michael Howard, David LeBlanc and John Viega. “24 Deadly Sins of Software Security”. “Sin 5: Buffer Overruns.” Page 89. McGraw-Hill. 2010.

[REF-7] Mark Dowd, John McDonald and Justin Schuh. “The Art of Software Security Assessment”. Chapter 6, “Signed Integer Boundaries”, Page 220. 1st Edition. Addison Wesley. 2006.

Content History

Submissions			
Submission Date	**Submitter**	**Organization**	**Source**
	CLASP		Externally Mined
Modifications			
Modification Date	**Modifier**	**Organization**	**Source**
2008-09-08	CWE Content Team	MITRE	Internal
	updated Applicable_Platforms, Background_Details, Common_Consequences, Relationships, Relationship_Notes, Taxonomy_Mappings, Weakness_Ordinalities		
2008-11-24	CWE Content Team	MITRE	Internal
	updated Relationships, Taxonomy_Mappings		
2009-10-29	CWE Content Team	MITRE	Internal
	updated Common_Consequences, Relationships		
2010-12-13	CWE Content Team	MITRE	Internal
	updated Background_Details		
2011-06-01	CWE Content Team	MITRE	Internal
	updated Common_Consequences		
2011-09-13	CWE Content Team	MITRE	Internal
	updated Relationships, Taxonomy_Mappings		
2012-05-11	CWE Content Team	MITRE	Internal
	updated Common_Consequences, Demonstrative_Examples, References, Relationships		
2014-07-30	CWE Content Team	MITRE	Internal
	updated Relationships, Taxonomy_Mappings		

2 Comments on CWE-128 Description

2.1 Description and Context

CWE-128 is weakness base that means a weakness that is described in an abstract fashion, but with sufficient details to infer specific methods for detection and prevention. It is more general than a Variant weakness, but more specific than a Class weakness. The context of CWE-128, using UML class diagram, is given in Fig. 1.

Categories are higher level weaknesses:

- CWE-189: Numeric Errors – Weaknesses in this category are related to improper calculation or conversion of numbers.
- CWE-742: CERT C Secure Coding Section 08 - Memory Management (MEM) – Weaknesses in this category are related to rules in the memory management section of the CERT C Secure Coding Standard. Since not all rules map to specific weaknesses, this category may be incomplete.
- CWE-876: CERT C++ Secure Coding Section 08 - Memory Management (MEM) - Weaknesses in this category are related to rules in the Memory Management (MEM) section of the CERT C++ Secure Coding Standard. Since not all rules map to specific weaknesses, this category may be incomplete.
- CWE-998: SFP Secondary Cluster: Glitch in Computation – This category identifies Software Fault Patterns (SFPs) within the Glitch in Computation cluster.

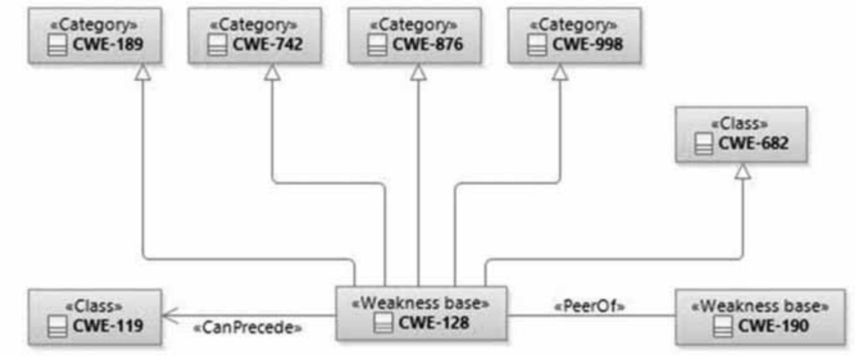


Fig. 1. CWE-128 Context

All these categories have simply classification purpose: “Category – A CWE entry that contains a set of other entries that share a common characteristic.”

Quality of CWEs classification is not subject of this paper.

CWE-682: Incorrect Calculation – “The software performs a calculation that generates incorrect or unintended results that are later used in security-critical decisions or resource management.

When software performs a security-critical calculation incorrectly, it might lead to incorrect resource allocations, incorrect privilege assignments, or failed comparisons among other things. Many of the direct results of an incorrect calculation can lead to even larger problems such as failed protection mechanisms or even arbitrary code execution."

CWE-682 is a weakness class, i.e. "A weakness that is described in a very abstract fashion, typically independent of any specific language or technology. It is more general than a Base weakness." This means that CWE-682 is abstract one and has classification purpose.

Above mentioned categories and class add some clarification for CWE-128 but in essence nothing concrete enough to be used for this weakness detection.

One more element is the direct association CanPrecede with CWE-119 – Improper Restriction of Operations within the Bounds of a Memory Buffer. This association is in the sense that CWE-128 can be used before CWE-119 in attack CAPEC-92 – Forced Integer Overflow.

CWEs are only partially ordered with the relation CanPrecede. This association is used when it is detected from some well-known attack. CanPrecede is associated with concrete CAPEC.

Finally, CWE-128 is a peer of CWE-190: Integer Overflow or Wraparound. This means that CWE-128 is something like CWE-190:

"Description Summary: The software performs a calculation that can produce an integer overflow or wraparound, when the logic assumes that the resulting value will always be larger than the original value. This can introduce other weaknesses when the calculation is used for resource management or execution control.

Extended Description: An integer overflow or wraparound occurs when an integer value is incremented to a value that is too large to store in the associated representation. When this occurs, the value may wrap to become a very small or negative number. While this may be intended behavior in circumstances that rely on wrapping, it can have security consequences if the wrap is unexpected. This is especially the case if the integer overflow can be triggered using user-supplied inputs. This becomes security-critical when the result is used to control looping, make a security decision, or determine the offset or size in behaviors such as memory allocation, copying, concatenation, etc.

Terminology Notes: "Integer overflow" is sometimes used to cover several types of errors, including signedness errors, or buffer overflows that involve manipulation of integer data types instead of characters. Part of the confusion results from the fact that 0xffffffff is -1 in a signed context. Other confusion also arises because of the role that integer overflows have in chains."

CWE-128 is not bounded only to integers as CWE-190. In that sense CWE-128 is more abstract than CWE-190, i.e. CWE-128 can exist for floats and other types.

On the other hand, CWE-128 is bounded only with wraparounds, but not overflows. In that sense, now, CWE-190 is more abstract than CWE-128 because it is dealing with overflows too.

At least, wraparound for integers is common for CWE-119 and CWE-190. The context of last one is somehow similar to that of CWE-119.

Classifications of CWEs, CVEs and CAPECs are not exclusive ones. For example, CWEs have common properties as is mentioned above for CWE-119 and CWE-190.

CWEs, CVEs and CAPECs are connected together, but these links are only partly included in their descriptions.

Conclusion from this investigation is that more efforts are needed to achieve well-structured databases for CWEs, CVEs and CAPECs.

2.2 Focus on CWE-128

This weakness is applicable for C and C++. Therefore CWE-128 must be investigated for these 2 programming languages.

CWE-190 deals with integer overflows or wraparounds. In CWE-190 integer overflow is used as synonym of wraparound.

CWE-128 deals with wraparounds. Its description contains an example with integers, but nowhere is mentioned something about the data type for which this weakness is applicable. Theoretically speaking, CWE-128 is applicable to float types too, but pragmatically, it is CWE-190 for C/C++ - at least for now.

More arguments for above conclusion that CWE-128 is CWE-190 for C/C++ are:

- Both CWEs have no children of any weakness kind.
- CWE-128 is mapped to a subset of the nodes to which CWE-190 is mapped in the other classifications. The only difference is that of CLASP. The arguments for this difference are not clear, but pragmatically they can be ignored using above mentioned reasons.

So, conclusion is that CWE-128 is dealing with integer types in C/C++ and “wraparound” is a synonym of “integer overflow”.

The next question is “Which are integer (integral) types in C/C++?” The answer is: int with variants signed/unsigned, short, and long. Here are bool and char (signed/unsigned). A special case of integers are enum that are sets of integers. Pointers and references are integers too. There are no reference arithmetic, therefore references can be excluded from further investigations.

CWE-128 is applicable “…whenever a value is incremented past the maximum value for its type…”. What exactly means “incremented”? Wraparound does not occurs only when operation increment is applied. The list of all operations is ++,

--, +, -, *, >>, <<, &, ^, |, ++=, --=, +=, -=, *=, >>=, <<=, &=, ^= and |=.

All these operations are applicable to all above mentioned types, because, at least, the operands are implicitly converted to suitable integer type.

The compiler can generate code for these operations and in debug mode check results from operations, but in non-debug mode it is possible CWE-128 to occur.

Another approach is the source code to be verified for possibilities CWE-128 to occur and the developer to put needed operand checks as preconditions. Verification is better approach because it is possible CWE-128 never to occur for some operations depending of the calculation context.

There is no integer overflow exception in C/C++, so, the programmer must take actions to elude it.

3 Conclusion

After this thorough analysis of CWE-128 it is time to formalize above mentioned findings.

References

1. CWE-128: Wrap-around Error, http://cwe.mitre.org/data/definitions/128.html
2. Vladimir Dimitrov, CWE-119 in Z-Notation, Proc. of Ninth International Conference ISGT’2015, 2016, pp. 90-94.

Detailed Formal Specification of Software Weakness CWE-128 (Specification)

Vladimir Dimitrov,

Faculty of Mathematics and Informatics, University of Sofia, 5 James Bourchier Blvd.,
1164 Sofia, Bulgaria
cht@fmi.uni-sofia.bg

Abstract. Software weaknesses are described in formatted text. There is no widely accepted formal notation for that purpose. This paper shows how Z-notation can be used for formal specification of CWE-128. The weakness is specified at different detail levels and informal level environment are discussed.

Keywords: software weakness, formalization, Z-notation.

1 Formalization of CWE-128 in Z-notation

"Why Z-notation?" is motivated in details in [2].

First, types and their limits must be defined:

$bool == \mathbb{Z}$

$char == \mathbb{Z}$

$signed\_char == \mathbb{Z}$

$unsigned\_char == \mathbb{N}$

$wchar\_t == \mathbb{Z}$

$char16\_t == \mathbb{Z}$

$char32\_t == \mathbb{Z}$

$short == \mathbb{Z}$

$signed\_short == short$

$unsigned\_short == \mathbb{N}$

$int == \mathbb{Z}$

$signed\_int == int$

$unsigned\_int == \mathbb{N}$

$long == \mathbb{Z}$

$signed\_long == long$

$unsigned\_long == \mathbb{Z}$

$long\_long == \mathbb{Z}$

$signed\_long\_long == long\_long$

$unsigned\_long\_long == \mathbb{N}$

$enums == \mathbb{F}_1\ \mathbb{Z}$

V. Dimitrov, V. Georgiev (Editors): ISGT'2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

Here enums and pointers are defined but they are analyzed later on.

$$\begin{array}{|l}
\mathit{CHAR\_MIN}, \mathit{CHAR\_MAX}, \mathit{SCHAR\_MIN}, \mathit{SCHAR\_MAX}\colon \mathbb{Z} \\
\mathit{UCHAR\_MAX}\colon \mathbb{N}_1 \\
\mathit{WCHAR\_MIN}, \mathit{WCHAR\_MAX}\colon \mathbb{Z} \\
\mathit{UINT\_LEAST}16\mathit{\_MAX}, \mathit{UINT\_LEAST}32\mathit{\_MAX}\colon \mathbb{N}_1 \\
\mathit{SHRT\_MIN}, \mathit{SHRT\_MAX}\colon \mathbb{Z} \\
\mathit{USHRT\_MAX}\colon \mathbb{N}_1 \\
\mathit{INT\_MIN}, \mathit{INT\_MAX}\colon \mathbb{Z} \\
\mathit{UINT\_MAX}\colon \mathbb{N}_1 \\
\mathit{LONG\_MIN}, \mathit{LONG\_MAX}\colon \mathbb{Z} \\
\mathit{ULONG\_MAX}\colon \mathbb{N}_1 \\
\mathit{LLONG\_MIN}, \mathit{LLONG\_MAX}\colon \mathbb{Z} \\
\mathit{ULLONG\_MAX}\colon \mathbb{N}_1 \\
\hline
\mathit{CHAR\_MIN} < \mathit{CHAR\_MAX} \wedge (\forall c\colon \mathit{char} \bullet \mathit{CHAR\_MIN} \leqslant c \leqslant \mathit{CHAR\_MAX}) \\
\mathit{SCHAR\_MIN} < \mathit{SCHAR\_MAX} \wedge (\forall c\colon \mathit{signed\_char} \bullet \mathit{SCHAR\_MIN} \leqslant c \leqslant \mathit{SCHAR\_MAX}) \\
(\mathit{CHAR\_MIN} = \mathit{SCHAR\_MIN} \vee \mathit{CHAR\_MIN} = 0) \\
(\mathit{CHAR\_MAX} = \mathit{SCHAR\_MAX} \vee \mathit{CHAR\_MIN} = \mathit{UCHAR\_MAX}) \\
(\forall c\colon \mathit{unsigned\_char} \bullet c \leqslant \mathit{UCHAR\_MAX}) \\
\mathit{WCHAR\_MIN} < \mathit{WCHAR\_MAX} \wedge (\forall c\colon \mathit{wchar\_t} \bullet \mathit{WCHAR\_MIN} \leqslant c \leqslant \mathit{WCHAR\_MAX}) \\
(\forall c\colon \mathit{char}16\mathit{\_t} \bullet c \leqslant \mathit{UINT\_LEAST}16\mathit{\_MAX}) \\
(\forall c\colon \mathit{char}32\mathit{\_t} \bullet c \leqslant \mathit{UINT\_LEAST}32\mathit{\_MAX}) \\
\mathit{UINT\_LEAST}16\mathit{\_MAX} \leqslant \mathit{UINT\_LEAST}32\mathit{\_MAX} \\
\mathit{SHRT\_MIN} < \mathit{SHRT\_MAX} \wedge (\forall i\colon \mathit{short} \bullet \mathit{SHRT\_MIN} \leqslant i \leqslant \mathit{SHRT\_MAX}) \\
\mathit{INT\_MIN} < \mathit{INT\_MAX} \wedge (\forall i\colon \mathit{int} \bullet \mathit{INT\_MIN} \leqslant i \leqslant \mathit{INT\_MAX}) \\
\mathit{LONG\_MIN} < \mathit{LONG\_MAX} \wedge (\forall i\colon \mathit{long} \bullet \mathit{LONG\_MIN} \leqslant i \leqslant \mathit{LONG\_MAX}) \\
\mathit{LLONG\_MIN} < \mathit{LLONG\_MAX} \wedge (\forall i\colon \mathit{long\_long} \bullet \mathit{LLONG\_MIN} \leqslant i \leqslant \mathit{LLONG\_MAX}) \\
(\forall i\colon \mathit{unsigned\_long\_long} \bullet i \leqslant \mathit{ULLONG\_MAX}) \\
\mathit{LLONG\_MIN} \leqslant \mathit{LONG\_MIN} \leqslant \mathit{INT\_MIN} \leqslant \mathit{SHRT\_MIN} \\
\mathit{SHRT\_MAX} \leqslant \mathit{INT\_MAX} \leqslant \mathit{LONG\_MAX} \leqslant \mathit{LLONG\_MAX} \\
\mathit{SCHAR\_MAX} \leqslant \mathit{UCHAR\_MAX} \\
\mathit{SHRT\_MAX} \leqslant \mathit{USHRT\_MAX} \\
\mathit{INT\_MAX} \leqslant \mathit{UINT\_MAX} \\
\mathit{LLONG\_MAX} \leqslant \mathit{ULLONG\_MAX}
\end{array}$$

These above mentioned relations among limits are defined but still generic. Concrete values must be set for particular implementation, like:

$$\begin{array}{|l}
\mathit{SCHAR\_MIN}, \mathit{SCHAR\_MAX}\colon \mathbb{N} \\
\ldots \\
\hline
\mathit{SCHAR\_MIN} = -127 \\
\mathit{SCHAR\_MAX} = 127 \\
\ldots
\end{array}$$

As defined by the standard, applicable operations for the type bool are &, |, ^, ~ and ++ (in postfix and prefix version). Following the standard, these operations could not generate integer overflow in any case. The operation ++ in any case generate only true value.

It is possible, implicitly or explicitly, some value to be converted to another type and integer overflow for the new type to be available, but the specifications here are dealing only with the original type.

How to react when an integer overflow occurs?

$$EXCEPTION ::= No \mid IntegerOverflow$$

In case of integer overflow, an exception is raised and the result is undefined. There is no need the value of undefined result to be further analyzed – for example to be set of undefined results, the exception is raised and what exactly is the value of the result is not important. It is enough to set a special single value to the result in that case.

$$\begin{array}{|l} BadResult\colon \mathbb{Z} \end{array}$$

Integers in Z-notations are infinite set and therefore integer overflow never occurs. So, if the result of calculation is out of limits of its type integer overflow must be set for this operation.

The operation addition for chars with check for integer overflow is defined as:

$$\begin{array}{|l}
\_AddChar\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \\
x?, y?\colon char \\
r!\colon char \\
ex!\colon EXCEPTION \\
\hline
CHAR\_MIN \leqslant x? + y? \leqslant CHAR\_MAX \land r! = x? + y? \land ex! = No \lor \\
(x? + y? < CHAR\_MIN \lor CHAR\_MAX < x? + y?) \land r! = BadResult \land ex! = IntegerOverflow \\
\hline
\end{array}$$

In the same way all operations for all types can be specified, but let try to make specifications more compressed.

The basic schema is:

$$\begin{array}{|l}
\_AddChar0\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \\
x?, y?\colon char \\
r!\colon char \\
\hline
r! = x? + y? \\
\hline
\end{array}$$

Internally, the result of the operation can be saved in a variable:

$$\begin{array}{|l}
\_AddChar1\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \\
x?, y?\colon char \\
r\colon char \\
r!\colon char \\
\hline
r = x? + y? \land r! = r \\
\hline
\end{array}$$

Then the schema with integer overflow check look like:

$\_AddChar2$

$x?, y?: char$

$r: char$

$r!: char$

$ex!: EXCEPTION$

$r = x? + y?$

$(CHAR\_MIN \leqslant r \leqslant CHAR\_MAX \wedge r! = r \wedge ex! = No \vee$

$(r < CHAR\_MIN \vee CHAR\_MAX < r) \wedge r! = BadResult \wedge ex! = IntegerOverflow)$

Now is possible to extract a schema for integer overflow check:

$\_Raised$

$r: char$

$r!: char$

$ex!: EXCEPTION$

$(r < CHAR\_MIN \vee CHAR\_MAX < r) \wedge r! = BadResult \wedge ex! = IntegerOverflow$

This schema could not be used alone because r has no value.
The next schema describes the ok case:

$\_NotRaised$

$r: char$

$r!: char$

$ex!: EXCEPTION$

$CHAR\_MIN \leqslant r \leqslant CHAR\_MAX \wedge r! = r \wedge ex! = No$

This schema, as the previous one, could not be used alone because r has no value.
Now the new schema with checks is:

$\_AddChar$

$x?, y?: char$

$r: char$

$r!: char$

$ex!: EXCEPTION$

$r = x? + y? \wedge (NotRaised \vee Raised)$

The variable r is working one and can be hidden from the schema interface in that way:

$\_AddChar$

$x?, y?: char$

$r!: char$

$ex!: EXCEPTION$

$\exists r: char \bullet r = x? + y? \wedge (NotRaised \vee Raised)$

Now, above presented schemas can be abstracted from the limits. The base of integral types are integer numbers:

$$\begin{array}{|l}
\multicolumn{1}{l}{\_Raised\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_} \\
r, MIN, MAX: \mathbb{Z} \\
r!: \mathbb{Z} \\
ex!: EXCEPTION \\
\hline
(r < MIN \vee MAX < r) \wedge r! = BadResult \wedge ex! = IntegerOverflow \\
\hline
\end{array}$$

$$\begin{array}{|l}
\multicolumn{1}{l}{\_NotRaised\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_} \\
r, MIN, MAX: \mathbb{Z} \\
r!: \mathbb{Z} \\
ex!: EXCEPTION \\
\hline
MIN \leqslant r \leqslant MAX \wedge r! = r \wedge ex! = No \\
\hline
\end{array}$$

$$\begin{array}{|l}
\multicolumn{1}{l}{\_AddChar\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_} \\
x?, y?: char \\
r!: char \\
ex!: EXCEPTION \\
\hline
\exists r, MIN, MAX: char \bullet r = x? + y? \wedge \\
MIN = CHAR\_MIN \wedge MAX = CHAR\_MAX \wedge (NotRaised \vee Raised) \\
\hline
\end{array}$$

In just a same way subtraction for chars can be defined:

$$\begin{array}{|l}
\multicolumn{1}{l}{\_SubChar\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_} \\
x?, y?: char \\
r!: char \\
ex!: EXCEPTION \\
\hline
\exists r, MIN, MAX: char \bullet r = x? - y? \wedge \\
MIN = CHAR\_MIN \wedge MAX = CHAR\_MAX \wedge (NotRaised \vee Raised) \\
\hline
\end{array}$$

The same approach is applicable for all other operations for chars.

Arithmetic in all integral types is based on the integer's arithmetic. The difference among the operations on integral types are the concrete limits of the type. The schemas for normal and abnormal executions are based only on the integers.

In the same way, the operations for all other integral types are defined, for example addition for unsigned short is:

$$\begin{array}{|l}
\multicolumn{1}{l}{\_AddUnsignedShort\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_} \\
x?, y?: unsigned\_short \\
r!: unsigned\_short \\
ex!: EXCEPTION \\
\hline
\exists r, MIN, MAX: char \bullet r = x? + y? \wedge \\
MIN = 0 \wedge MAX = USHRT\_MAX \wedge (NotRaised \vee Raised) \\
\hline
\end{array}$$

In CWE-128 is important the result. Hence, the type of input parameters need not to be specified – the hypothesis is that needed conversation (implicit or explicit) is already done. So, the leading example operation looks like:

AddChar

$x?, y?: \mathbb{Z}$

$r!: char$

$ex!: EXCEPTION$

$\exists r, MIN, MAX: \mathbb{Z} \bullet r = x? + y? \land$

$MIN = CHAR\_MIN \land MAX = CHAR\_MAX \land (NotRaised \lor Raised)$

Focus of CWE-128 is on the result now.

The arithmetic operations can have as an operand of type unenclosed enumeration but this is simply a named constant of integer type.

An operation with enum as lvalue is:

$E: enums$

AddEnum

$x?, y?: \mathbb{Z}$

$r!: \mathbb{Z}$

$ex!: EXCEPTION$

$\exists r: \mathbb{Z} \bullet r = x? + y? \land$

$(r \notin E \land r! = BadResult \land ex! = IntegerOverflow \lor$

$r \in E \land r! = r \land ex! = No)$

Here, E is set of integers (enum). This specification can be further augmented with the corresponding concrete values of the set as:

$E: enums$

$E = \{1, 3, 5\}$

If the result is not from the set then the exception is raised. Strictly speaking, this is not the case of CWE-128 because the result can be put in the placed enum place.

Finally is the pointer's case. An integer can be added or subtracted from the pointer and two pointers of the same type can be subtracted.

Actually, the pointer is an address – a natural number in the address space. Every pointer has a type size. The pointer type is:

$MAX\_ADDRESS: \mathbb{N}_1$

Pointer

$address: \mathbb{N}$

$type\_size: \mathbb{N}_1$

$address \leqslant MAX\_ADDRESS$

The basic pointer operations are:

*AddPointer*0

$\Delta Pointer$

$i?: \mathbb{Z}$

$address' = address + type\_size * i?$

*SubPointer*0

$\Delta Pointer$

$i?: \mathbb{Z}$

$address' = address - type\_size * i?$

SubPointers

$p1?, p2?: Pointer$

$r!: \mathbb{N}$

$r! =$

$(\mathbf{if}\ p1?.address \geqslant p2?.address$

$\mathbf{then}\ (p1?.address - p2?.address)$

$\mathbf{else}\ (p2?.address - p1?.address))\ \mathrm{div}\ p1?.type\_size$

Integer overflow is possible only in the first and the second operation.

AddPointer

$\Delta Pointer$

$i?: \mathbb{Z}$

$ex!: EXCEPTION$

$\exists r: \mathbb{Z} \bullet r = address + type\_size * i? \wedge$

$(0 \leqslant r \leqslant MAX\_ADDRESS \wedge address' = r \wedge ex! = No \vee$

$(r < 0 \vee MAX\_ADDRESS < r) \wedge ex! = IntegerOverflow \wedge address' = BadResult) \wedge$

$type\_size' = type\_size$

SubPointer

$\Delta Pointer$

$i?: \mathbb{Z}$

$ex!: EXCEPTION$

$\exists r: \mathbb{Z} \bullet r = address - type\_size * i? \wedge$

$(0 \leqslant r \leqslant MAX\_ADDRESS \wedge address' = r \wedge ex! = No \vee$

$(r < 0 \vee MAX\_ADDRESS < r) \wedge ex! = IntegerOverflow \wedge address' = BadResult) \wedge$

$type\_size' = type\_size$

2 Conclusion

Practically, CWE-128 can occur when some calculation is performed and then the result is assigned to some variable of integral, enumeration or pointer type. If the result does not fit in the memory location of this variable, then CWE-128 occurs. It is possible, the operation to be applied to different kind of operands with implicit and explicit conversion, but the result is important. If the result fits (in range) it is ok, else CWE-128 occurs.

How these specifications can be used? There are two approaches: white and black box.

In the first case only program interface is available to the tester. So, using the parameter semantics and program logic, the tester can put such test values to force CWE-128.

In the second case, program code is available. One approach is the full code review – in every place where above mentioned operations are used, the reviewer must investigate is it possible a value outside the result type limits to be generated, but it too time consuming. Second one is the code verification. In that case, assignment operators to variables from above mentioned types must be verified for post conditions that satisfy type limits.

References

1. CWE-128: Wrap-around Error, http://cwe.mitre.org/data/definitions/128.html
2. Vladimir Dimitrov, CWE-119 in Z-Notation, Proc. of Ninth International Conference ISGT'2015, 2016, pp. 90-94.

Integrating software applications for analysis of big heterogeneous data from parallel sequencing in a distributed cloud environment

Yana Nikolova[1], Vladimir Dimitrov[1], Luca Pireddu[2], Dimitar Vassilev*[1]

[1] Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”, 5 James Bourchier Str, Sofia 1164, Bulgaria

[2] CRS4, Polaris, Loc. Piscina Manna Ed. 1, 09010 Pula, ItalyUniversity of Cagliari, 09124 Cagliari, Italy

* Corresponding author: dimitar.vassilev@fmi.uni-sofia.bg

Abstract. The paper presents the integration of software tools for the analysis of big heterogeneous data from parallel sequencing in a cloud environment. The major applications of distributed analysis in the cloud environment are defined through Hadoop and MapReduce. The SEAL software was used for some of the processing steps, such as alignment of parallel sequencing data from a maize genome. The major methodological goal is to present an opportunity to develop a procedure to unite all the analysis steps in a pipeline. The particular bioinformatics application was focused on the implementation of Burrows-Wheeler Transformation approach for alignment of the short reads of the publicly accessed maize genome. The results emphasize the benefits stemming from the use of cloud computing infrastructure and distributed computing technologies to analyze sequencing data.

Keywords: Bioinformatics, parallel sequencing, big heterogeneous data analysis, cloud infrastructure, Hadoop, MapReduce

1 Introduction

With the continuous improvement in high-throughput sequencing technologies, the amount of sequencing data to process is growing very rapidly, sometimes faster than the computing resources. The speed of new bioinformatics data generation establishes a number of preconditions for the successful storage and analysis of the data – analysis that includes preprocessing, mapping, variant calling, annotation, and visualization [1,2,10,12,17]. Due to the complexities and falling data acquisition costs, the main budget voices in large genome sequencing studies are shifting from sequencing to data analysis, amounting of up to 80% [19]. Therefore, the main task of bioinformatics is defined as analyzing large datasets generated by high-throughput sequencing technologies. The data is sometimes heterogeneous, generated by the sequencers in various formats, and accompanied

V. Dimitrov, V. Georgiev (Editors): ISGT’2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

by various types of metadata relating to the technology of sequencing. In addition, for the purposes of the particular application all the sequencing data is often used together with other data (clinical records or other phenotypic data), which deepens their heterogeneous nature.

Working with heterogeneous data suggests the development and implementation of adequate methods and software tools for their integration, with the goal of achieving semantic interoperability of the relevant systems. Particularly important to a successful "cloudification" is a scalable integration of the nucleotide sequence alignment step, as it is one of the most resource- and time-intensive steps in the process. [5,9,15,16].

2 Theoretical Basics

The term "Big Data" defines the concept of data so large or complex that traditional applications for processing data are insufficient. Challenges include analyzing, collecting, searching, storing, sharing and visualization of data. An important part of understanding the concept of big data is the realization that the volume considered as "big" grows with time, as the computing resources become faster and cheaper. Regardless of the elastic nature of the concept of Big Data, there are basic characteristics that identify it: ***Volume*** (large amounts of data), ***Velocity*** (high rate of production or modification of the data), ***Variety*** (variety of types and sources of data) and ***Veracity*** (growing need to assess the veracity of the data with the speed of their generation). Analysis of Big Data includes the use of advanced techniques capable of handling very large datasets of different types and sizes – structured, unstructured, streaming, etc. This type of data cannot fit predefined models directly and it needs to be preprocessed first.

Generating such large volumes of data bears challenges. The data generation rate of parallel DNA sequencing technologies has been growing quickly in past several years [8,12,13,14]. This growth forces requirements on the implementation of new models for working with data [3,10,11,13]. Storage and archiving of the resulting data is expensive, so most laboratories prefer to re-sequence DNA instead of storing it.

As the technology of sequencing instruments is progressing, storage requirements for the downstream analysis are increasing. Although the instrument's expected output can be easily calculate based on the size of the genome being studied and the type of study – which imposes a coverage and a number of samples to achieve sufficient statistical significance – the storage requirements for the secondary analysis are far more difficult to predict, considering the variety of data types and different data analysis software. Secondary analysis includes several tasks that are both CPU and data intensive, such as mapping short reads to a reference genome, removal of duplicate alignments, and so on (Figure 1).

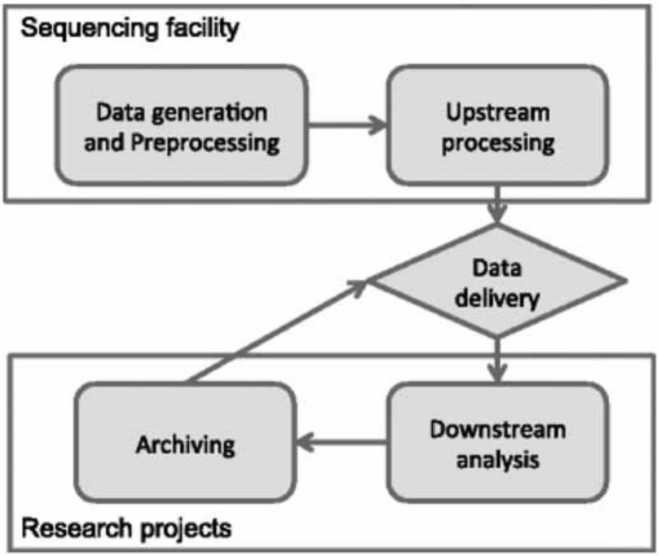


Figure 1. Upstream vs. downstream analysis (image from Spjuth et al.[15])

3 Data formats

The data, generated by next generation sequencing contains a lot of additional information such as metadata about the method of sequencing, encoding, clinical information and others. In order to be easily used, the data should be organized and normalized in specific file formats commonly used by the community.

The FASTA and FASTQ file formats for sequence data are the most commonly used formats for sequences and base quality data. They are text-based formats, with independent records, each carrying a sequencing data and metadata. The SAM (Sequence Alignment/Map) file format is another text-based format that stores sequencing data along with reference mapping information; the format is structures as one record per line and the individual fields are separated by tabs. The files start with a header. BAM files are a binary compressed version of SAM files, expressing the same data but more compactly. Moreover, BAM files can be indexed by alignment coordinate, allowing quick access to sequences any part of the file.

Finally, the VCF *(Variant Call Format)* format is a tab-delimited format for storing variant calls and individual genotypes [7]. It is able to store all variant types, from single nucleotide variants to large-scale insertions and deletions. The format is highly flexible and can store a wide variety of information regarding the individual genomic variants. It has already been adopted by a number of large-scale projects and is supported by an increasing number of software tools.

4 Cloud environment

All data generated by NGS methods can be defined as big data, since their size entails significant complexities in storing and analyzing the data. Although many sequencing laboratories have opted to deploy powerful computing resources, the bursty nature of the sequencing data production process makes for a strong argument for renting large amounts of cloud-based computing power for short periods of time. In such situations, resources become available and are paid for as needed. Several vendors operate large data centers, where they have deployed large computing and data storage resources and are able to rent them at competitive prices. Virtualization is a technology what allows a physical resource, like a server, to be subdivided and shared among several independent virtual servers – each using part of the available physical resources. Users of cloud resources only pay for the resources they use and this way they avoid the risk of overpaying on unneeded computing resources. In many cases, it can be cheaper to use cloud resources than to purchase and maintain local computing resources. [5, 16].

Hadoop is an open source collection of software for distributed processing of large data sets in a cluster of computing nodes. It is available under the Apache license. Programs on this framework operate in an environment that provides distributed storage and processing within the cluster. Hadoop is designed to be scalable up to thousands of machines – each of which provides local computing and storage resources.

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data in-parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks. Typically the compute nodes and the storage nodes are the same, because this configuration allows the framework to effectively schedule tasks on the nodes where data is already present, thus bringing the computation to the data and producing very high aggregate bandwidth across the cluster. The key innovation in MapReduce is…. Thus, a MapReduce program is structured in two parts – map and reduce (Figure 2). The map phase includes filtering and sorting the input data, and each node receives a copy of the map function, which performs locally on the data part hold in it. The reduce phase is an aggregation of the output of the map phase over values with the same key. Each node receives a copy of the reduce function which is executed and generates a new set of key-value pairs.

While widely applicable, the MapReduce paradigm is not well suited for a wide variety of computations. With YARN system included in version 0.23, Hadoop have management system for allocation of global resources for multiple applications, as well as their customization. YARN (Yet Another Resource Negotiator) is a software rewrite that decouples MapReduce's resource management and scheduling capabilities from the data processing component, enabling Hadoop to support more varied

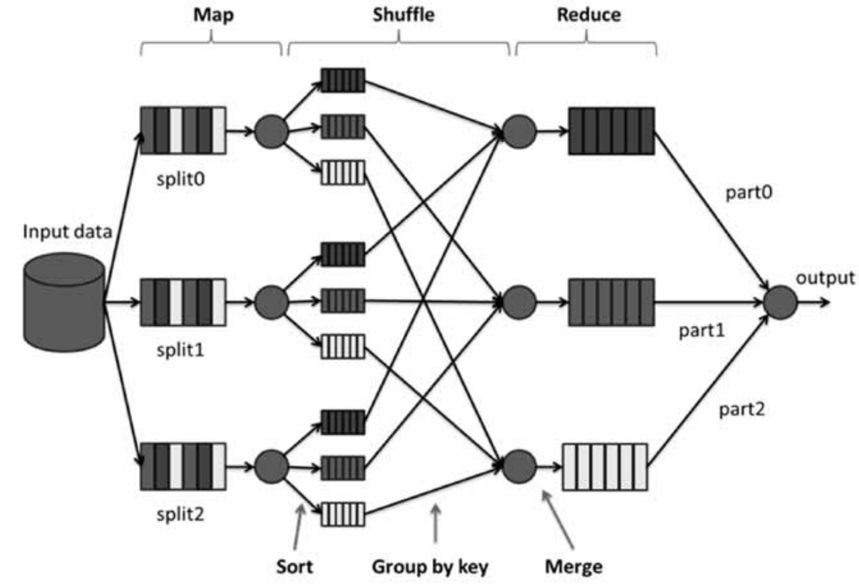


Figure 2. MapReduce pipeline

processing approaches and a broader array of applications. The fundamental idea is to split up the two major functionalities of the JobTracker, resource management and job scheduling and monitoring, into separate daemons. This is done by creating ResourceManager and per-application ApplicationMaster. An application is either a single job, in the classical sense of MapReduce jobs, or a set of jobs. The ResourceManager and per-node slave, the NodeManager, form the data-computation framework where the ResourceManager is the ultimate authority that arbitrates resources among all the applications in the system. The per-application ApplicationMaster is, in effect, an application-specific process that is tasked with negotiating resources from the ResourceManager and working with the NodeManager(s) to execute and monitor the tasks. Running applications in YARN consists in 3 main steps: sending a request for resources; loading the application in ApplicationMaster instance; executing the tasks.

The Hadoop framework also shows its own limitations when it is used for big data storage and analysis. Some of them are related to the security and suitability of the platform. The security model in Hadoop is disabled by default due to sheer complexity. Encryption at the storage and network levels is also missing. The framework is written almost entirely in Java. Due to its high capacity design, the HDFS (Hadoop Distributed File System) lacks the ability to efficiently support the random reading of small files and the platform is not suited for small data needs.

Although Hadoop is the most popular framework for batch processing of big data, there are several new framework other than Hadoop that are gaining popularity. Apache Spark promises faster speeds than Hadoop MapReduce along with good application programming interface. It can run in-memory on a cluster (if there is sufficient memory available) and it does not impose the two-stage MapReduce paradigm. Cluster MapReduce provides a Hadoop-like framework for MapReduce jobs run in a distributed environment, where by simplifying movement of data and minimizing dependencies that can show data pull, the performance is much faster. A massive parallel-processing platform, High Performance Computing Cluster (HPCC) incorporates a data refinery cluster called Thor, a query cluster called Roxie, plus middleware components, external communications and client interface. Hydra is a distributed task processing system, which can handle some of the big data tasks and supports streaming and batch options using tree-based data structure so it can store and process data across huge clusters.

Regardless the alternative technologies, MapReduce is created to work with big and heterogeneous data and it is still widely used for batch processing.

5 Methods for analyzing heterogeneous data in bioinformatics

The massive amount of biomedical data generated in recent years has posed some crucial problems stemming not only from the data's volume but also its specific features. Apart from some big projects there is a strong demand in developing new methods for analyzing big heterogeneous sequencing data.

The data used in bioinformatics is Big Data - it has big volume, it is unstructured and heterogeneous. Defined as such, it brings the challenges of storing and analyzing the data properly, using the new technologies.

The following steps describe the process of analyzing sequencing data:

- DNA sequencing – a real laboratory sequencing using any of the available techniques.
- Conversion of the sequenced data into different data formats needed for the used analysis software.
- Quality rating (scoring) – the raw data representing sequences of the four letters of nucleotide bases (A, C, G, T) in combination with letter N when the base cannot be determined. They are checked for errors in determining the base, low quality reads.
- Alignment (mapping) of reads to a reference genome (alignment) of these sequences to create the copy of the original nucleotide sequence.
- Variant calling: finding different types of mutations, insertions, deletions and other structural and non-structural variations with respect to the reference genome.

- Functional annotation of the variants.

After the qualitative assessment, the data is formed by millions and sometimes billions of short reads from unknown genomic positions. Unfortunately, there are no sequencing technologies that can read the entire DNA sequences, or even parts of considerable length. This requires the use of algorithms for assembling. The major steps in analyzing sequencing data are: assembly, mapping, variant calling and annotating.

Alignment is an optimization problem in which gaps are inserted gradually between the sequences in order to minimize the distance between them. The aim is sequences to be approximated as much as possible with less embedded intervals. The alignment can be of two types – global or local. Algorithms for global alignment try to impose two sequences throughout their length, while locally alignment algorithms try to find similar subsequences on which they focus. Generally all algorithms for alignment fall into three categories – hash based, methods using the "seed-and-extend" paradigm and algorithms based on string transformation. The transformation algorithms used in bioinformatics are numerous, but in general the most widely used are the Needleman-Wunsch method, the Smith-Waterman method and the Burrows-Wheeler method. [1,6]

5.1 Burrows-Wheeler transformation

The Burrows-Wheeler transformation (BWT) algorithm is used in bioinformatics from the beginning of 2000. The Burrows-Wheeler method is a transformation without losses, and the original string can be easily retrieved, because the transformation only switches the places of the characters in it (table 1).

Table 1. Burrows-Wheeler transformation example of the string "CTAGAGAA"

Transformation				
Input	All Rotations	Sorting All Rows into Lex Order	Taking Last Column	Output Last Column
CTAGAGAA	CTAGAGAA	AGAGAACT	AGAGAAC**T**	TGGCAAAA
	ACTAGAGA	AGAACTAG	AGAACTA**G**	
	AACTAGAG	AACTAGAG	AACTAGA**G**	
	GAACTAGA	TAGAGAAC	TAGAGAA**C**	
	AGAACTAG	GAGAACTA	GAGAACT**A**	
	GAGAACTA	GAACTAGA	GAACTAG**A**	
	AGAGAACT	CTAGAGAA	CTAGAGA**A**	
	TAGAGAAC	ACTAGAGA	ACTAGAG**A**	

The trie (comes from re**trie**val) for reference string is a data structure that allows fast string operations. It contains all possible substrings (prefixes for prefix trie or suffixes for suffix trie) of a reference string. Each string character is stored on a node (edge) with each substring being delimited with special character - $ marks the end of substring in suffix trie and ^ marks the start of substring in prefix trie. In a prefix trie, concatenating all characters on the nodes from a lead to the root forms a unique substring. Because exact repeats of the reference string are located under the same path of a trie, the query string needs to be matched (aligned) to only one copy of the repetitive region. The method can be easily parallelized by considering every substring matching as separate task.

There are many software implementations of the Burrows-Wheeler transformation. Below are listed some of the most commonly used software for analysis of NGS data.

SOAP is a software package for efficient comparison of short oligonucleotide reads generated by parallel sequencing to a reference genome. Besides detecting errors, indels and empty bases, the software achieves maximum performance by integrating the use of parallelism, multithreaded and GPU-based dynamic programming. GATK is software framework with rich set of tools for different types of analysis of the NGS data, including modeling errors, comparison of data and options for pre-programming code to the needs of use. Bowtie / Bowtie2 is primarily used in comparing the huge amount of short reads with large reference genome where for indexation of the reference genome is used the BWT method. BWA is used in comparison with similar short reads with large reference genome. It provides an effective method for complex calculations by breaking them down into smaller tasks and gives the best results when comparing two sequences, detecting matches, mismatches and the gaps between them.

Seal is a suite of distributed applications for aligning short DNA reads, and manipulating and analyzing short read alignments [4]. The software provides wide set of Hadoop-based tools for NGS data analysis. Seal applications generally run on the Hadoop framework and are made to scale well in the amount of computing nodes available and the amount of the data to process. This fact makes Seal particularly well-suited for processing large data sets. Thanks to its scalability, Seal can achieve very high throughput using the maximum computing resources available. With the ability to use shared memory, 7 or 8 alignment processes can be run concurrently on a single workstation with 8 GB of memory, using a human reference genome. Thanks to Hadoop, the alignment processes are monitored and controlled by Hadoop’s daemons, providing, automatic monitoring and restart of failed tasks. This way Seal provides a start-and-forget solution, resisting node failures and transient cluster conditions that would otherwise cause jobs to fail. It also avoids basing all operations on a centralized shared stored volume, which can represent a single point of failure and a performance bottleneck.

Because Seal is designed to run on Hadoop and use its distributed computing functionality, it is an effective choice to deal with the growing requirements of the big and heterogeneous data in bioinformatics.

6 Software realization

There exist a number of software packages for short read alignment, but unfortunately most of them are written following a conventional computing paradigm and thus need to be integrated into a distributed or Cloud environment to effectively respond the increasing demands of genomic data. To build an example of a new advanced solution to the problem, it is necessary to find the right distributed platform for Cloud environment, method for analyzing bioinformatics data and software to implement that method (figure 3).

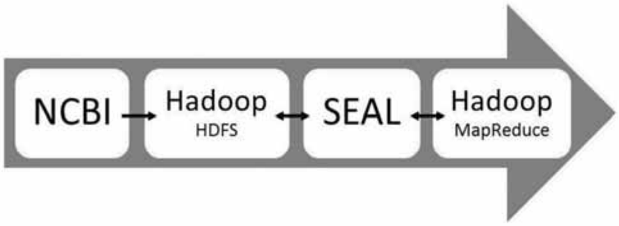


Figure 3. Workflow

Hadoop was chosen as the framework on which to build an effective processing pipeline because it is greatly simplifies the parallel execution of Big Data tasks in a distributed environment. The Burrows-Wheeler transformation is one of the most widely used methods for nucleotide sequences mapping to a reference genome. It is used by a number of software packages.

Upon selection of software for use in bioinformatics, it is import to consider a number of specific characteristics. The Seal package was selected as a typical example of a very good integration of bioinformatics software for short reads alignment in a distributed environment using the Hadoop framework. Seal is a necessary transitional step in the process of integrating existing bioinformatics software and methods in the innovative concept of distributed and Cloud environments.

For the purpose of the current work, a Linux machine was configured with the following characteristics: Debian OS v7.9, 64-bit; 146 GB storage; 28 GB RAM; 24 CPUs. Also, a number of auxiliary software was installed as required by Hadoop and Seal: Java 1.7; Python 2.7.3; Hadoop-BAM 7.1.0; **Hadoop 2.7.2**; **Seal (develop)**.

The tests in this project were performed using maize genome sample. Maize is the most commercial plant in the world and best sequenced plant genome available for use by the National Center of Biotechnology Information (NCBI). The data is available for public download or direct use and analysis through the website of NCBI. Part of the reason why the maize genome was seleted for this study is the fact that most software for analyzing genomic data primarily assume human or animal genomes, since they receive the most attention.

The data was downloaded from NCBI in FASTQ format, but because of specification in the input data format for Seal, the raw data is additionally parsed. A parser was created to add specific symbols to the header of every file – “tab /1” for first part of the paired reads and “tab /2” for the second part. After that the parsed files are loaded onto HDFS.

The first step in the analysis is to run the “seal prq” command, which reads all FASTQ files as input from HDFS directory, processes them using Hadoop MapReduce functionality to join paired reads that may be in different files into the same record, and finally generates files in PRQ format on HDFS.

The next step is to download a reference genome to which the short reads will be aligned. The reference genome of maize is downloaded from “MaizeGDB” public site and is indexed with the BWT. Indexing the reference genome generates several additional files, which are collected in a tar archive and loaded onto HDFS to be distributed by Hadoop onto all the cluster nodes, where the Seal aligner can access them during the process of mapping to the reference genome.

In this final step, the “seal align” tool receives as an input the PRQ files with the short reads of the nucleotide sequences and the archived file with the indexed reference genome of the maize. By using Hadoop MapReduce functionality, the tool maps the reads to the most likely location on the reference genome from where they were originally cut. The output file containing the aligned reads is stored back onto HDFS.

7 Conclusion

The use of distributed cloud system for processing and storing heterogeneous parallel sequencing big data is a new approach that can meet the ever increasing demand in different implementations.

The opportunities of integration and parallelization of processing and storage in distributed Cloud environment also gives big expectations for implementation of various algorithmic procedures and solutions for heterogeneous data analysis comprising the major algorithms and stages of the whole processing pipeline by achieving faster speed and lower cost.

Cloud services are way to achieve these objectives and although there are a number of concerns regarding their usability, they promise to have wide avenue for development and applications in bioinformatics, not only because they provide almost unlimited computing capabilities, but also because a specialized bioinformatics clouds can be build and available to all researchers

Acknowledgements: The presented work has been funded by the Sofia University SRF within the “Methods for Automated Semantic Annotation of Heterogeneous Datasets” Project, Contract No. 15/2016.

References

1. Green CS, Tan J, Ung M, Moore JH, Cheng C (2014). Big Data Bioinformatics. Wiley Periodicals. 229(12):1896-900. doi: 10.1002/jcp.24662.
2. Brown S M (2013). Next-Generation DNA Sequencing informatics. Cold Spring Harbor Laboratory Press. ISBN 978-1-936113-87-3.
3. Misra S, Agrawal A, Liao W K, Choudhary A (2011). Anatomy of a hash-based long read sequence mapping algorithm for next generation DNA sequencing. Bioinformatics (2011) 27 (2): 189-195. doi:10.1093/bioinformatics/btq648.
4. Pireddu L, Leo S, Zanetti G (2011). SEAL: a distributed short read mapping and duplicate removal tool. Bioinformatics, 27(15), 2159–2160.
5. Rijmenam M (2015). Big Data Hadoop Alternatives: What They Offer and Who Uses Them. (https://datafloq.com/read/Big-Data-Hadoop-Alternatives/1135).
6. Крачунов М (2014). Изкуствен интелект в биоинформатиката: автоматизиран анализ и класификация на данни от паралелно секвениране. Докторска дисертация. ФМИ на СУ „Св. Климент Охридски“.
7. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A. (2011) 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics, 27(15), 2156–2158.
8. Spujth O., Bongcam-Rudloff E., Guimera R.V., Kallio A., Korpelainen E., Kanduła M., Krachunov M., Kreil D., Kulev O., Łabaj P., Lampa S., Pireddu L., Schönherr S., Siretskiy A., Vassilev D. (2015) Experiences with workflows for automating data-intensive bioinformatics. *Biology Direct, 10:43 doi:10.1186/s13062-015-0071-8*.
9. O'Driscol, A., Daugelaite, J., Sleator R.D. (2014) 'Big data', Hadoop and cloud computing in genomics. Journal of Biomedical Informatics 46: 774–781
10. Bao R. et al. (2014) Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. *Cancer Informatics* 13(s2) 67–82 doi: 10.4137/CIn.s13779.
11. Lapatas V., Michalis Stefanidakis M., Jimenez R., Via A., and Schneider M.V. (2015) Data integration in biological research: an overview. *Journal of Biological Research-Thessaloniki* (2015) 22:9 DOI 10.1186/s40709-015-0032-5
12. Viari A. (2012) 'Big Data' en biologie. Médecine/Sciences, 28 : 1027-1028
13. Binder H., Blettner M. (2015) 'Big Data' in medical sciences – a biostatistical vew. *Deutsches Ärzteblatt International | Dtsch Arztebl Int*, 112: 137–42
14. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195
15. Spjuth O., Bongcam-Rudloff E., Dahlberg J., Martin D., Kallio A., Pireddu L., Vezzi F., Korpelainen E. (2016). Recommendations on e-infrastructures for next-generation sequencing. GigaScience 5:26. DOI**:** 10.1186/s13742-016-0132-7
16. Stein L. The case for cloud computing in genome informatics. Genome Biol. 2010; 11:207. doi:10.1186/gb-2010-11-5-207
17. Marx V. (2013) The big challenges of 'Big Data'. Nature 498:255-260
18. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195
19. Sboner A., Jasmine Mu, Greenbaum D., Raymond K Auerbach R.K., Mark B Gerstein M.B. et al.: The real cost of sequencing: higher than you think! Genome Biology 2011, 12:125, doi:10.1186/gb-2011-12-8-125

The necessity of Project Management capacities and Organizational Flexibility for the successful implementation of complex projects in Municipalities

Ioannis Patias

Faculty of Mathematics and Informatics
University of Sofia St.Kliment Ohridski“
5 James Bourchier blvd., 1164, Sofia, Bulgaria
ioannis.patias@gmail.com

Abstract. This paper presents the case of a team, successfully completed an initially planned for two and a half years complex system integration project, in just nine months, for a mid-sized Municipality. It was quite challenging complex system integration project in the sphere of Intelligent Transport Systems (ITS). The project was co-financed by the European Union (EU), and this added to the complexity due to the numerous regulations. Generalizing the findings the author provides two critical success factors (CSFs) for project implementation. On the one side it is the Project Management (PM) capacities and on the other the Organizational Flexibility (OF), that both affect the successful implementation of complex integration projects in Municipalities. The results generalization aims to provide a practical guide useful for similar structures. The focus should be on strengthening PM capacities, but at the same time try to develop OF in such levels, which can secure projects’ success.

Keywords: Project Management (PM), project complexity, Organizational Flexibility (OF), critical success factors (CSFs), Intelligent Transport Systems (ITS).

1 Introduction

The actuality of the problem is given; having in mind that in Bulgaria the transition of the municipalities from operations management based structures to project management has just begun within the framework period 2007-2013. Before that there was no really need for project management orientation of Municipality as a structure. The main focus was until recent on ensuring that business operations continue efficiently by using the optimum resources needed and meeting customer (in the concrete case the citizens) demands Thus, no matter the project management organizational maturity on behalf of the contractor, again the lack of capacity remains on the Municipality’s side.

V. Dimitrov, V. Georgiev (Editors): ISGT'2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

2 Literature Review

In this paragraph some theoretical background terms, and definitions are provided.

2.1 Project Management

According to PMI's A Guide to the Project Management Body of Knowledge (PMBOK® Guide)[1], project is a temporary endeavor undertaken to create a unique product, service or result. A project is temporary in that it has a defined beginning and end in time, and therefore defined scope and resources. And a project is unique in that it is not a routine operation, but a specific set of operations designed to accomplish a singular goal. So a project team often includes people who don't usually work together – sometimes from different organizations and across multiple geographies. Based on that, Project Management (PM), is the application of knowledge, skills, tools, and techniques to project activities to meet the project requirements.

On the other side Operations Management is an area of management concerned with ongoing production of goods and/or services. It involves ensuring that business operations continue efficiently by using the optimum resources needed and meeting customer demands. It is concerned with managing processes that transform inputs (e.g., materials, components, energy, and labor) into outputs (e.g., products, goods, and/or services).

2.2 Project Complexity

Complexity has different meanings for different people and in different organizations [2][3]. But, what is important is how organizations anticipate, comprehend and navigate complexity, and by that we can determine their successes or failures.

The most common characteristics of complex projects are [4]:

- Multiple stakeholders
- Ambiguity of project features, resources, phases, etc.
- Significant political/authority influences
- Unknown project features, resources, phases, etc.
- Dynamic (changing) project governance
- Significant external influences

2.3 Organizational Flexibility

Volberda's model on organizational flexibility [5] addresses how the companies should manage their dynamic capabilities and organizational design, in order to achieve the desired fit by being flexible. He studied how the organizations deal with the paradox of flexibility over time, which means, how they continuously adapt to the changes in the environment and balance corporate discipline with entrepreneurial creativity. He develops a strategic flexibility framework to configure the resources of the firm for effective responses to organizational change providing a comprehensive set of variables and their linear relationships. In addition to this argument, Volberda anticipated the possibility of modelling the adaptation process from a dynamic point of view: "Flexibility is not a static condition, but it is a dynamic process. Time is a very essential factor of organizational flexibility".

Volberda also defines dimensions of Extensiveness of flexibility mix in three levels Limited, Medium, and Broad as:

- Limited: The firm has developed a large ability to change the volume and mix of business activities. The firm dominates the operational flexibility "routine manoeuvring capacity",
- Medium: The firm has a good level of operational flexibility but a greater ability to change the organization structure and decision-making and communication processes. That is, the firm dominates the structural flexibility "adaptive manoeuvring capacity", and
- Broad: The firm has a good level of operational and structural flexibility but a greater ability to change corporate strategy and the nature of business activities. That is, the firm dominates the strategic flexibility "strategic manoeuvring capacity".

2.4 CSFs

Critical success factors (CSFs), also known as Key Results Areas (KRAs), refer to the activities that must be completed to a high standard of quality in order to achieve the goals of your project. CSFs are a way to prioritize certain tasks as the project plan is being executed. Having clear CSFs helps the project team clarify what needs to be worked on first or needs special attention, allowing people to work together to achieve the project's main objectives [6].

3 Case study

The requirements for the system the project was to deliver, covering the needs of the Municipality, were quite heavy (see figure 1).

An Automated Ticketing System, which should provide:

- module for storage and processing of data allowing the collection and analysis of data from the underlying transactions, and setting the types and prices of transport documents.
- central processing system of ticket sales and card meeting all the requirements for registration and reporting of fiscal devices, according to Bulgarian legislation and in particular the Road Transport Act and Regulation N18 for

registration and reporting of sales outlets by fiscal devices.

- proceeds from sales reported to the National Revenue Agency.
- system allows for downloading the database of transactions carried out by on-board ticketing systems in buses, portable readers and machines Ticket stops and strategic places in the city.
- all transactions transferred to the central system without any losses. During the transmission data will be encrypted.   Data transfer will be implemented through GRPS, WiFi, as well as topically.
- system data stored and archived on disk arrays included in the hardware equipment in the control center.
- system installed at the central level provides an opportunity for statements of transactions and revenue generated by each service.
- functionalities related to the provision of statistical information and analysis facilitates the management and the monitoring of the transport network.

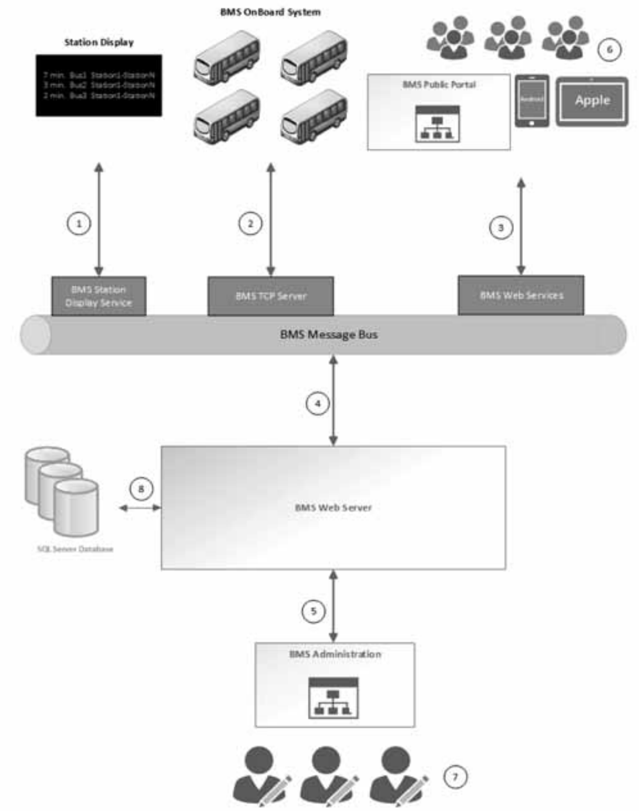


Figure 1: General view of architecture

- BI module, which contains a wide range of with tools to generate reports and manage all data stored in the database, including information on the number of passengers, passenger flow, as well as an analysis of each dynamic information reflected in the base.
- module for revenue management and resources is fully integrated with the other modules of the system.
- integration is a level data functionality and user interface.
- web - based system and must be accessed from each of the workstations via web browser
- central database system stores all data generated by all devices that are part of the system
- data must be transferred to the central system in the Control Center, where after the return of bus depot, information on underlying transactions during the day to be transferred remotely.
- 4 points with WiFi access and access to the system from depot

A system for ensuring the advantage of mass urban transport vehicles at intersections, and informs the passengers in real time, which should provide:

Advantage of mass urban transport vehicles at intersections by the following main components:

- bus equipment - Each of Bus Rapid Transit (BRT) corridor buses equipped with the corresponding transmitter.
- IDs approaching a crossroads - Traffic light controllers at intersections of BRT corridor upgraded with the necessary transceivers elements (two in each intersection) providing the ability to require local priority of vehicles. In the case of radio link, it will be encrypted to eliminate false claims priority.
- modernization of signaling modules - The location of the controllers at traffic lights that will require modifications in order to ensure priority passage of buses at intersections, and a provisional number of points to give signal priority passage for each location in accordance with the approved investment project based on the selected decision on the route of the BRT corridor. The point at which the alert transfer it to the relevant controller, as described above.
- CCTV cameras (CCTV), which will be installed on certain major intersections in BRT corridor - installed cameras at intersections

System for passenger information in real time including the following components:

- dispatching equipment - work places
- communication module controller-driver – all vehicles equipped with communication module dispatcher guide with touch screen display indicating the current status of movement of buses for your route according to the schedule

- equipment for vehicles - The system provides tools with which board system bus generates a pulse to account for the bus communicate with the central system.
- information boards at bus stops and strategic places in the city. All stops serving fast bus lines and other strategic locations in the city equipped with information boards with LED displays mounted outside the bus-shelters
- information boards in vehicles
- other information interfaces to: website, mobile application notification via SMS
- information boards and interfaces receive online information in real time
- Automatic vehicle location (AVL) system

4 Discussion - results generalization

Apart the above technical complexity described, "bottlenecks" were generated in all project phases, which we can generalize as having the following causes:

- Ambiguity of project features, resources, phases, etc. together with lack of administrative capacity on behalf of the Municipality, at the level of two people with no previous experience of such complex integration project implementation were responsible,
- Lack of authority to the appointed PM on behalf of the Municipality. The project team on behalf of the contractor had to face a typical operations oriented structure, thus whenever something was appearing and needed a decision, either precious time was wasted until the corresponding department answers, or the issue had to go up to the Mayor for resolution, creating additional significant political/authority influences and a dynamic (changing) project governance and
- Multiple stakeholders combined with resistance to new. All the involved stakeholders were limited informed for their responsibilities. On the central level key documents, like the new urban transport regulations, were not in place, thus everybody was avoiding to undertake responsibilities, not officially described and delegated to him, having in mind the importance of the project and the significant external influences.

All the above created an unsecure environment for the project. The project team on behalf of the contractor had to both fight with those uncertainties and fast track the project implementation. What was used to help resolve the uncertainties and deliver on-time the project:

- Project Management (PM) capacities

  PM capacities were invested in terms of supporting the Municipality PM team in performing its tasks. The project team on behalf of the contractor invested time and efforts and the result was to establish an environment of trust. After all, following the procedures is a stable instrument for any situation of lack of secureness.

- Organizational Flexibility (OF)
  The Municipality PM team having this feeling of trust, was then transformed into a flexible component of an overall mix structure (according to the described above). In other words the Municipality, initially Limited was transformed into Broad Extensiveness of flexibility mix and respectively structure.

The result was great. The project was delivered on-time, on-budget and responding to the required quality. The gains were much more than limited in the project itself. The Municipality PM team at the end of the project had the confidence of further implementing even more complex projects.

The applied stakeholder management was focused on trust and engagement in order to accelerate project's implementation. It was not hierarchical, or organizational focused but purely on the project's tasks in a consistent and organized manner, which finally contributed to project success.

Having in mind the changing nature of the stakeholders' commitment to the project and the relationships between different Departments and Municipality structures as the project progresses the Contractor's project team focused broadly not only on their own role as stakeholder. The focus was on all key stakeholders in the Municipality and how they influence the project, and this helped the both the Contractor and Municipality teams to find the way to collaborate.

5 Conclusions

The typical operations management model is under transformation in the Municipalities and it is about to be changed by a more project-oriented model. What needs to be done on behalf of the contractors is focus on two critical success factors (CSFs) for project implementation. On the one side it is the Project Management (PM) capacities both on behalf of the Contractor and on behalf of the Municipality. Whatever is missing puts the project in risk. On the other side the Organizational Flexibility (OF) affects the successful implementation of complex integration projects in Municipalities. We have always to bear in mind that the Municipalities are not keen in working on a project-oriented approach, and try to balance. The focus should be on strengthening PM capacities, but at the same time try to develop OF in such levels, which can secure projects' success.

References

1. Project Management Institute (Corporate Author), A Guide to the Project Management Body of Knowledge 5th Edition, ISBN-13: 978-1933890517 (PMI, 2013)
2. Evgeniy Krastev, Maria Semerdjieva,"Business Process Model Based on Business Rules", Proceedings of the 8th Int. Conference Information Systems & GRID Technologies 2014
3. Evgeniy Krastev, Kristiyan Shahinyan, Computer Assisted Quality Assessment of a Set of Business Process Models, Proceedings of the 9th IEEE European Modelling Symposium of Mathematical Modelling and Computer Simulation, IEEE Computer Society, 2015, pp.180-186, doi:10.1109/EMS.2015.36
4. Project Management Institute, Navigating Complexity: A Practice Guide (PMI, 2014)
5. Volberda, H.W., Building the Flexible Firm, Oxford: Oxford University Press, ISBN-13: 978-0198295952, (1998).
6. Spalek, S, Critical Success Factors In Project Management. To Fail Or Not To Fail, That Is The Question!, http://www.pmi.org/learning/critical-success-factors-project-management-7568, PMI, 2005

Management of heterogeneous data collections from external sources through PostgreSQL FDW

Svetoslav Savov*[1], Vladimir Dimitrov[1], Dimitar Vassilev[2, 1]

[1] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", 5 James Bourchier Str, Sofia 1164, Bulgaria
[2] Bioinformatics group, AgroBioInstitute, 8 Dragan Tsankov Blvd, Sofia 1164, Bulgaria
* Corresponding author: svetlio_81@yahoo.com

Abstract. In the modern world huge web projects use different programming languages for separate parts of the software. This polyglot approach is also applicable for the databases. While the Relational Database Management System (RDBMS) is common it has its limitations when it comes to managing highly variable data. Database engines' developers have provided alternative solutions based on NoSQL databases. One of the most popular open-source RDBMS is PostgreSQL. It follows the SQL (Structured Query Language) rules and is famous for its stability, security, large community, many contributors and well documented functionalities. In order to meet the new tendencies and support the polyglot model (utilizing the advantages of different data sources) in the projects' development the PostgreSQL programmers have introduced a way to query data from external sources through FDW (Foreign Data Wrappers). FDW in PostgreSQL allow effective data management of many external sources of information, including analysis of heterogeneous data collections.

Keywords: heterogenous big data, PostgreSQL, NoSQL, FDW, MongoDB

1 Introduction

FDWs are based on the SQL Management of External Data (SQL/MED) standard which supports the SQL access to remote data sources and objects through corresponding interfaces. The feature has been officially supported after PostgreSQL 9.1. The full list of the released FDW can be found at the PostgreSQL wiki page. FDW are drivers that allow the PostgreSQL database administrators to run queries and get data from many external sources like other SQL databases (Oracle, MySQL), NoSQL databases(MongoDB, Redis, CouchDB), text files in CSV and JSON formats, content from Twitter and many more. Some of the wrappers allow both reading and writing operations on the remote data to be completed through PostgreSQL (figure 1). The current development strategies are focused on combining the advantages of more software solutions in order to achieve better performance and security for the projects.

Heterogeneous data collections contain information with different structure that should be gathered, stored and analyzed in a compound system. Such data sets

V. Dimitrov, V. Georgiev (Editors): ISGT'2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

are typical for contemporary applications that manage voluminous structured and unstructured data from different sources. It is usually recognized as Big Data. In order to extract the necessarily knowledge the heterogeneous big data collections should properly queried and the results should be analyzed. While the information from homogenous data sets can be easily described with predefined structure and managed through a relational database system, the management of heterogeneous data is more complicated. The reason is that we can not use the same structure for data from different sources and nature.

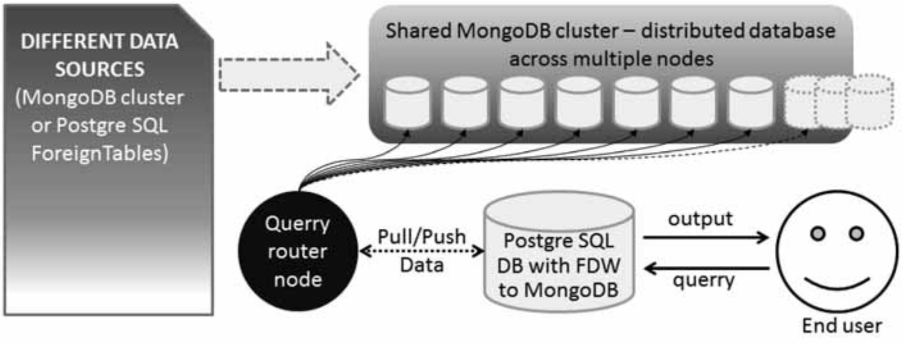

DIFFERENT DATA SOURCES (MongoDB cluster or Postgre SQL ForeignTables)
Shared MongoDB cluster – distributed database across multiple nodes
Querry router node
Pull/Push Data
Postgre SQL DB with FDW to MongoDB
output
querry
End user

Figure 1. Polyglot persistence model

Usually the heterogeneous big data collections are stored in NoSQL databases like MongoDB. It stores the data in JSON schema-less documents. Instead of flat two-dimensional tables typical for the relational databases they could contain heterogeneous big data collections of recursive, multi-dimensional objects. While MongoDB is valued for its easy scalability like the rest of the NoSQL database technologies, one of the reasons for its success is its usage as a data structure store that allows the programmers to have the same programing model in both the code and the database. The last will simplify the application's development and avoid the complex mapping between the program's logic and the relational structure where the data is separated in many tables.

It is expected that MongoDB or a similar NoSQL database engine will become a preferred database solution for operational data storage with relational databases used as an additional specialized tool for complex queries.

2 Problems and technical approaches for the heterogeneous big data management

The main obstacles in gaining the expected profit from the heterogeneous big data collections one might face during the data analysis are the complexity, heterogeneity and the decentralization of the data sources. It is important to handle the huge volume of data within a timely manner, to analyze it, to properly determine the relevance and gain a certain value of the obtained knowledge. Nowadays most the organizations that work with Big Data still struggle to correctly manage and merge the information from heterogeneous data flows in order to come with informative decisions for their business.

In order to successfully battle with the described problems the organizations should have stable storage solutions, usually based on distributed Cloud architecture. Also, the applications that gather, store and work with the corresponding heterogeneous big data collections should maintain a very high level of security and information privacy. The data should be handled and fully analyzed in timely manner. Usually it is based on data mining strategies and algorithms for automation of the knowledge's extraction and recognition processes.

To meet the organizations' requirements for complex business strategies and processes programmers use advanced approaches and proper models that combine the strengths of different technologies. Such an example is the combination of the powerful query engine provided by one of the leading RDBMS - PostgreSQL and a popular document oriented NoSQL database - MongoDB. The foreign data wrapper that is used for the establishment of this communication allows the users to leverage the MongoDB benefits from within PostgreSQL (figure 2).

An example of such a compound system could be a portal that gathers sales reports and logs from many different small online retail shops and big e-commerce web sites that sale different goods and services. A generalized model will be shown in the described test case.

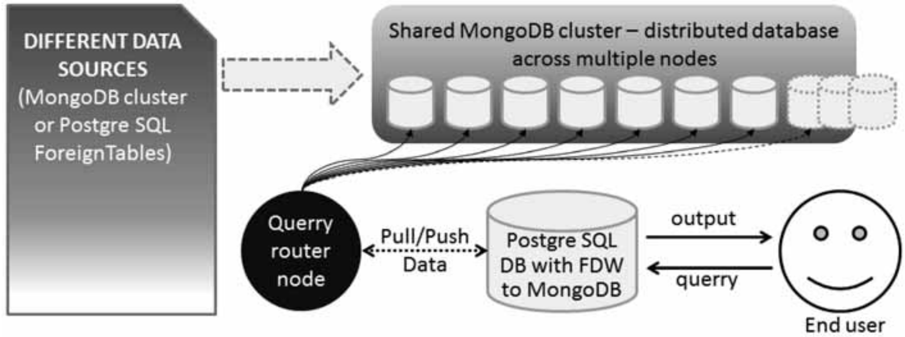

DIFFERENT DATA SOURCES (MongoDB cluster or Postgre SQL ForeignTables)
Shared MongoDB cluster – distributed database across multiple nodes
Querry router node
Pull/Push Data
Postgre SQL DB with FDW to MongoDB
output
querry
End user

Figure 2. PostgreSQL – MongoDB compound system

3 Strengths and weaknesses of Relational and NoSQL databases

Most of the web applications are using relation databases like MySQL and PostgreSQL. They have been developed for decades and they are perfectly suitable to store data which structure layout is known in advance. Two dimension tables which are suitable for the description of most business models can be created. The relation databases allow complex queries to be executed towards the data and content from different tables to be visualized through the provided JOIN operators. They are secure and flexible in the way the structured data is retrieved. The data is kept consistent.

On the other hand the RDBMS is not suitable when it is required to store data with huge variations in the records' structure and many hierarchical sub-levels. In these cases NoSQL database models which allow much more freedom on the data structure, simpler management, less system requirements, fast scalability on multiple servers and fast performance can be used. However, the consistency of the data is not always preserved. They allow the storage of multidimensional structures with huge amount of data.

The SQL database models follow the Atomicity, Consistency, Isolation and Durability (ACID) properties. They are used to store important data like financial transactions and accounts for example. Due to their nature and the lack of the same level of consistency NoSQL databases are used to store less important data like server logs or data that is quite variable and can not be easy described in a structure during the design stage.

In order to preserve the security level, utilize the numerous features of the PostgreSQL databases and benefit from the performance and scalability of the NoSQL models the contributors have developed FDW.

4 Connecting PostgreSQL with MongoDB through FDW

MongoDB is a widely used NoSQL solution. It is a document oriented database. It allows objects with different number of fields to be included in a database. It is a popular way to store Big Data from various sources. What is more, objects can be nested in other objects and there is no limit on the depth.

In order to set a working environment PostgreSQL should be installed and configured on the server. There are pre-built packages for most Linux distributions. For this study case get and install the correct PostgreSQL 9.2 rpm package for the server architecture by running commands similar to the following ones:

```
wget http://yum.postgresql.org/9.2/redhat/rhel-
6-i386/pgdg-centos92-9.2-6.noarch.rpm
rpm -ivH pgdg-centos92-9.2-6.noarch.rpm
```

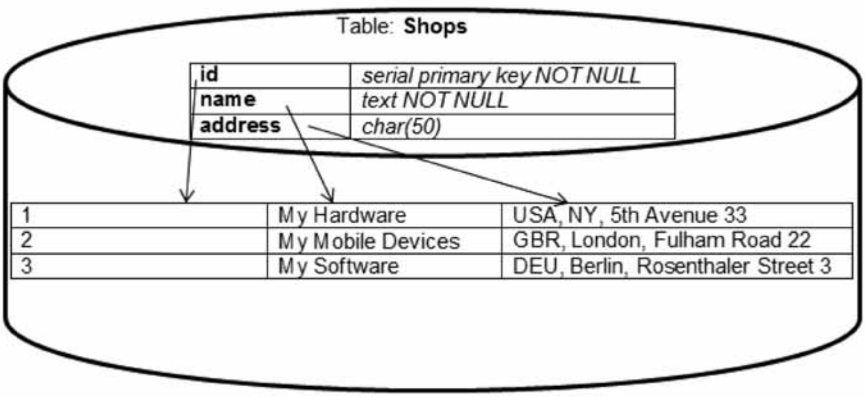
The `yum search postgres` command should be used to list all the available packages for the chosen server architecture. The packages can be installed with the following command:

```
yum install postgresql92-devel postgresql92-server
postgresql92-contrib
```

The PostgreSQL cluster should be initialized and started:

```
   service postgresql-9.2 initdb
   Initializing database:               [ OK ]
   /etc/init.d/postgresql-9.2 start
   Starting postgresql-9.2 service:          [ OK ]
```

Once the PostgreSQL database server is up and running create a test table with some sample data (figure 3):



id	*serial primary key NOT NULL*
name	*text NOT NULL*
address	*char(50)*

1	My Hardware	USA, NY, 5th Avenue 33
2	My Mobile Devices	GBR, London, Fulham Road 22
3	My Software	DEU, Berlin, Rosenthaler Street 3

Figure 3. Test table

In this example a system that collects the total income from different types of online shops can be created. Each shop might sell totally different products and it might be difficult to define the tables‘ structure during the design stage.

A document oriented database like MongoDB which better supports the storage of highly variable data can be used.

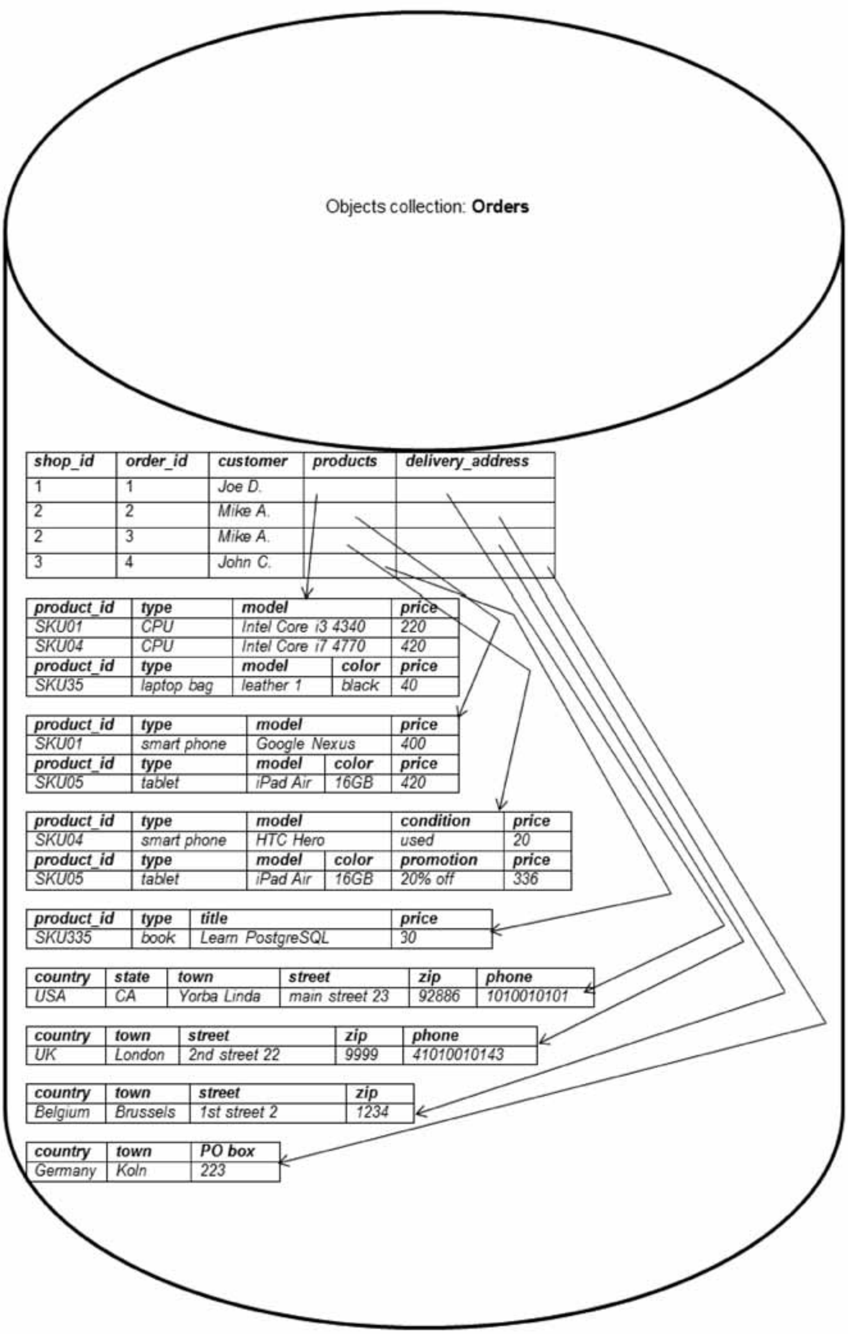
Create the corresponding `/etc/yum.repos.d/mongodb.repo` file with the configuration details for the MongoDB repository as explained in MongoDB’s official installation instructions. Then use the `yum install mongo-10gen mongo-10gen-server` command to install a stable release of the MongoDB server and the included tools. Start the service by entering `the /etc/init.d/mongod start` command in the command prompt.

Both PostgreSQL and MongoDB services can auto-start after the system is rebooted. This can be achieved by entering the `chkconfig postgresql-9.2 on && chkconfig mongod on` commands in the terminal.

The command shell can be started by typing `mongo` followed by the database name:

```
  mongo myshops
   MongoDB shell version: 2.4.8
   connecting to: myshops
```

Next, enter some sample objects in a MongoDB database (figure 4):



shop_id	order_id	customer	products	delivery_address
1	1	Joe D.		
2	2	Mike A.		
2	3	Mike A.		
3	4	John C.		

product_id	type	model	price
SKU01	CPU	Intel Core i3 4340	220
SKU04	CPU	Intel Core i7 4770	420

product_id	type	model	color	price
SKU35	laptop bag	leather 1	black	40

product_id	type	model	price
SKU01	smart phone	Google Nexus	400

product_id	type	model	color	price
SKU05	tablet	iPad Air	16GB	420

product_id	type	model	condition	price
SKU04	smart phone	HTC Hero	used	20

product_id	type	model	color	promotion	price
SKU05	tablet	iPad Air	16GB	20% off	336

product_id	type	title	price
SKU335	book	Learn PostgreSQL	30

country	state	town	street	zip	phone
USA	CA	Yorba Linda	main street 23	92886	1010010101

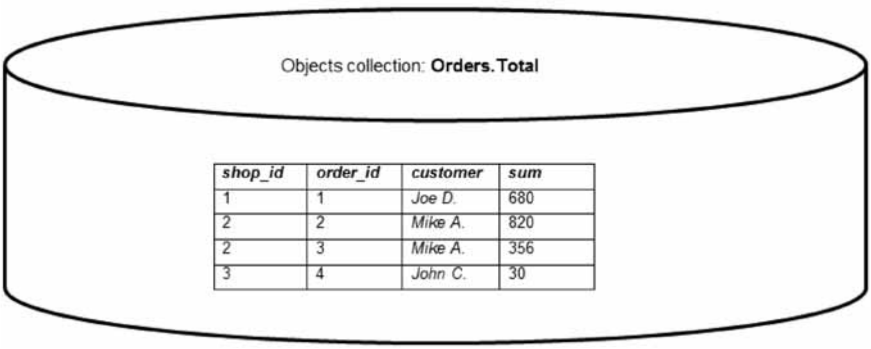
country	town	street	zip	phone
UK	London	2nd street 22	9999	41010010143

country	town	street	zip
Belgium	Brussels	1st street 2	1234

country	town	PO box
Germany	Koln	223

Figure 4.-- Sample objects

The built-in MongoDB aggregation functionality can be used to get the total sum for every order. It can be stored in a separate data collection. It will be needed later in the study case (figure 5).



shop_id	order_id	customer	sum
1	1	Joe D.	680
2	2	Mike A.	820
2	3	Mike A.	356
3	4	John C.	30

Figure 5. Case study

Once the sample data is stored in the PostgreSQL and MongoDB databases they are ready to be bound through the FDW. First, use git to get the driver from the repository. Then build and install the wrapper:

```
  cd /usr/src/
  git clone https://github.com/citusdata/mongo_fdw
  cd /usr/src/mongo_fdw/
  PATH=/usr/pgsql-9.2/bin/:$PATH make
  PATH=/usr/pgsql-9.2/bin/:$PATH make install
```

Next, load the extension from the PostgreSQL command line interface. Verify it and create a server instance for the wrapper:

```
postgres=# CREATE EXTENSION mongo_fdw;
CREATE EXTENSION
postgres=# \dx mongo_fdw;
       List of installed extensions
 Name | Version | Schema |     Description
------+-------+------+-----------------------------
 mongo_fdw | 1.0 | public | foreign data wrapper for
  MongoDB access
(1 row)
postgres=# CREATE SERVER mongo_server FOREIGN DATA
  WRAPPER mongo_fdw OPTIONS (address ‚127.0.0.1‘,
  port ‚27017‘);
CREATE SERVER
```

Proceed with the foreign table setup:

```
CREATE FOREIGN TABLE shops_sales
```

```
(
  shop_id INTEGER,
  order_id INTEGER,
    customer TEXT,
    sum INTEGER
)
SERVER mongo_server
OPTIONS (database ‚myshops‘, collection ‚Orders.
  Total‘);
```

It is ready to run SQL queries on the data originally stored in the MongoDB database. For example list all the records from the table. Then run another query to find the total income for each shop and sort the result based on the shop ID.

```
SELECT * FROM shops_sales;
 shop_id | order_id | customer | sum
---------+----------+----------+-----
       3 |        4 | John C.  |  30
       2 |        2 | Mike A.  | 820
       1 |        1 | Joe D.   | 680
       2 |        3 | Mike A.  | 356
(4 rows)

SELECT shops.id AS „shop ID“, shops.name AS „shop name“,
SUM(shops_sales.sum) AS „income“ FROM shops INNER JOIN
shops_sales ON shops.id = shops_sales.shop_id GROUP BY
shops.id ORDER BY shops.id;
 shop ID |     shop name     | income
---------+-------------------+--------
       1 | My Hardware       |    680
       2 | My Mobile Devices |   1176
       3 | My Software       |     30
(3 rows)
```

5 FDW future

FDW work as mediators between the PostgreSQL databases and numerous heterogeneous Big data sets stored in MongoDB collections. SQL queries can be executed on every possible source of information as long as the wrapper knows how to convert the external data to the PostgreSQL compatible format. It will extract the data from the underlying technology and unify the way to query it.

Disclaimer

The test case has been completed with MongoDB FDW (mongo_fdw 2.0.0), and PotsgreSQL 9.2 on a CentOS 6 machine. While the used commands might be differ for other software releases the approach should be similar.

Acknowledgments. The presented work has been funded by the Sofia University SRF within the "Methods for Automated Semantic Annotation of Heterogeneous Datasets" Project, Contract No. 15/2016.

References

1. http://wiki.postgresql.org
2. http://docs.mongodb.org
3. http://en.wikipedia.org/wiki/ACID
4. http://www.postgresql.org/docs/manuals/
5. https://wiki.postgresql.org/wiki/Foreign_data_wrappers
6. http://www.jamesserra.com/archive/2015/07/what-is-polyglot-persistence/
7. https://github.com/EnterpriseDB/mongo_fdw

Separation of Concerns in Motion Path Control of Redundant Robot arms

Evgeniy Krastev

Faculty of Mathematics and Informatics, St. Kliment Ohridski University of Sofia, 5 James Bourchier Blvd., 1164 Sofia, Bulgaria

eck@fmi.uni-sofia.bg

Abstract. This paper focuses on the application of an object- oriented approach for motion control of redundant robot arms along a path in task space. We demonstrate that blending methods in software engineering with mathematical methods can produce new solutions to existing problems in robotics. In the beginning we the problem of task planning from an object oriented viewpoint. Further on we apply the Separation of concerns principle to interpret the major types of concerns in an extended space of configurations of a redundant robot arm. This allows us to treat the robot arm motion in the extended space of configurations in a similar way the end- effector motion executes in task space. Further on, the selected approach enables a visual representation of the Null space of a redundant robot arm in the context of its motion. The visual interpretation of the Null space is used to solve the Jacobian singularity problem. This problem is of major concern for kinematic control of manipulators. An efficient numerical criterion is proposed to identify a Jacobian singularity. The Jacobian singularity problem is solved in terms of separating areas of concerns and applying the Principle of continuity

Keywords: separation of concerns, redundant robot arm, kinematics, motion path control, Jacobian computation, Jacobian singularity

1 Introduction

Robotics has become one of the most interdisciplinary fields for scientific research even before the invention of the first industrial robot Unimate in 1961. The list of core disciplines in robotics comprises mechanical engineering, electrical engineering and computer science. Until recently the successful application of robots in a wide spectrum of human activities has been attributed mainly to the advance of technologies related to mechanical and electrical engineering. Software technologies employed in industrial robots are less mature than the software technologies accepted as a standard in the software industry nowadays [1] [2] [3]. Most of the software for robot arm motion programming is based on proprietary programming languages and concepts typical for the 1970s. This kind of software applications are tightly coupled and therefore, it is difficult to reuse and maintain such applications. Even within

NASA they find it easier to design and build the software components for every new robot from scratch [4].

Therefore in the last few years we witness a growing interest to implement in robotics established standards, concepts and technologies for software development, such as the Rational Unified Process and Object- oriented, software engineering and object oriented programming. Robotics itself is well suited to accept object oriented technologies for software development [5]. Indeed, it is natural to view a robot as assembled of multiple bodies, representing objects with state and behavior, ready to execute commands and respond to events. The new trend encourages outlining systems and layers of functional subsystems and structuring them in a modern the software architectural style, where the implementation of software engineering patterns enable code reuse and easy adaptation to inevitable changes in technologies. An example case for system architecture of layered embedded components applied in a distributed robotic system is presented in [6] [7]. On the other hand, this new approach to software development allows to discover more efficient solutions to problems in robotics.

Most of the software applications in industrial robotics are related to motion control of robot arms. A robot arm executes a wide spectrum of activities in a factory environment. Repetitive tasks, such as welding, spraying, pick- and- place operations are usually pre- planned operations in task space. The task space dimension is determined by the number of parameters required to execute a given operation. For instance, a welding operation requires three parameters to describe the linear position and one parameter to describe angular position of the welding torch along the direction of the seam. It is common practice to preplan the work operation for a particular robot only. In reality, however, the operation preplanning is independent of the robot arm that is selected to execute it. Because the specific parameters of a given robot arm are tightly coupled in this software application it cannot be reused to execute the same operation with another robot arm. Moreover, it makes pre- planning of robot arm operations rather expensive and inflexible to changes in production plans.

This example briefly illustrates the advantage of applying an object oriented- oriented approach in a problem that is typical for robotics. The operation and the robot arm represent two separate entities in terms of object- oriented software engineering. The model of the work task will consider them as independent subsystems with well-defined interfaces allowing them to provide and consume services, as well as, to respond to events in the working environment [8].

The relation between the dimension $m, m \leq 6$ of the task space and the degrees of freedom n, $n \geq m$ of the robot arm selected to execute the work task has an important impact on the quality of the robot arm motion. In case, $n > m$ the robot arm has redundant degrees of freedom that can be used to satisfy motion quality criteria. Robot arms with redundant degrees of freedom are called redundant robot arms. By definition redundancy is task dependent. The motion of redundant robot arms is characterized by spatial flexibility of motion close to that of a human arm. In a variety of work tasks this is a major prerequisite to execute the task. Together with the ability to optimize the robot motion during the work task execution it makes redundant robot arms an appealing subject for research [9] [10] [11].

This paper presents a solution to the problem of Jacobian singularities in motion planning of redundant robot arms in terms of an object oriented approach. For many years this problem has been considered as the main obstacle for the wide spread application of Jacobian – based models for motion control along a given path in task space. Currently all known solutions aim avoiding singularities [12]. They use heavy numerical methods like Singular Value Decomposition [13], manipulability measure [14] or the condition number of the Jacobian to estimate the proximity to a singularity or how to estimate the proximity to a singularity again for the purpose of avoiding it [15].

The novelty in the proposed solution is the implementation of the Separation of concerns principle in software engineering to develop a mathematical model that solves the Jacobian singularity problem in motion control along a given path in task space. This object oriented approach identifies motion and geometry as two major types of concerns. The layer of motion concern is represented in an extended space of configurations. The application of vector space methods [16] allows to visualize the Null space of the redundant robot arm for arbitrarily selected point on the path in task space. The most important results are related to identifying and processing Jacobian singularities. The Continuity principle is employed to develop a heuristic procedure allowing to pass- through singularities instead of avoiding them.

The objective of this paper is to demonstrate the advantages in applying an object oriented approach for resolving the inverse kinematic problem for motion control of a redundant robot arm along a given path in task space. In the following section we formulate the problem of task planning from an object oriented viewpoint. In Section 3 the Separation of concerns principle is applied to represent the major types of concerns in an extended space of configurations. The new results in applying this approach are presented in Sections 4 and 5. Section 4 provides a visualization of the Null space of a redundant robot arm. The visual interpretation of the Null space is exploited in Section 5 to solve the Jacobian singularity problem. This problem is of major concern for kinematic control of manipulators. An efficient numerical criterion is proposed to identify a Jacobian singularity. The Jacobian singularity problem is solved in terms of separating areas of concerns and applying the Principle of continuity.

2 Problem statement

Let a robot arm manipulation task is defined in terms of a $m, m \leq 6$ dimensional space. The coordinates in task space define the set of parameters needed to describe the position of a characteristic point H and the angular orientation of the frame $Hxyz$ fixed in the tool employed to execute a work operation along the path γ. The position of H and the orientation of $Hxyz$ are described with respect to the base frame $O_o x_o y_o z_o$

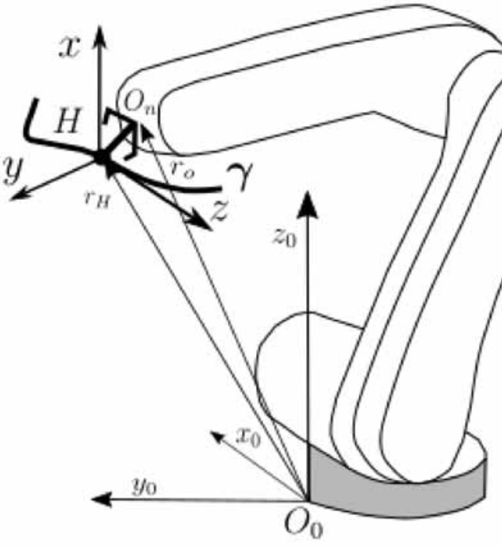


Fig. 1. Robot arm position in task space.

The orientation of the tool is fully determined by three parameters (Euler or roll-pitch and yaw representation). Without loss of generality, we consider the parameters for position and orientation as functions of an independent variable λ and use them to represent parametrically the path $\gamma = \gamma(\lambda)$ of the tool in task space. For clarity, we introduce the following parametrical representation for γ

$$\gamma : \Lambda \rightarrow R^m, \text{ where } \Lambda = [\lambda_1, \lambda_2]. \tag{1}$$

This way we specify the geometrical path of the tool in task space during the work operation (Fig. 1).

The description of the work operation would be incomplete, if we miss to describe the motion of the work tool along the path γ. For instance, the welding torch must reach the starting point of the welding seam at a predefined velocity. Thereafter the tool must move at constant velocity along the seam in order to distribute evenly the welding intensity. In practice the motion parameters are given by means of the velocity vector

$$\partial\gamma/_{\partial\lambda}(\lambda)\dot{\lambda} \tag{2}$$

The motion over γ is defined by the funciton

$$\lambda = \lambda(t), \mathrm{t} \in [t_0, t_1] \tag{3}$$

The expression (2) represents in a generalized form the velocity in terms of the coordinates of the linear velocity of the point H in the work tool and the coordinates of the vector of the angular velocity of the frame $Hxyz$ with respect to the base frame $O_o x_o y_o z_o$. From (1) it follows that $\partial\gamma/_{\partial\lambda}(\lambda)\dot{\lambda}$ is an m- dimensional vector.

The direction of motion over γ and the magnitude of the velocity are determined by the function of motion $\lambda(t)$.

It is noteworthy that the two groups of task parameters (1) and (2) are specified independently of the robot arm description. Note also that the motion parameters (2) are defined over the path γ in terms of the function of motion $\lambda(t)$. This represents another important difference with similar treatments of the motion path planning problem in the existing literature, where the velocity (2) is not related to the points of a geometrical path such as γ [13]⍰. Here we follow the natural requirement to bind the motion parameters to the points of γ in accordance with the technological requirements of the work operation.

The geometric path parameters and the function of motion represent two different types of concerns. These concerns are independent from each other and must be treated this way in the mathematical model for motion control. Moreover, the thus assigned work operation is not tightly related to a specific robot arm. It could be executed by any robot arm possessing the necessary technical parameters allowing its end-effector to execute the work operation.

Consider now an industrial robot arm M with n degrees of freedom. Denote by $Int\ Q_M$ the interior of the set $Q_M = \{ q = (q_1, \ldots, q_n): q_i \in [a_i, b_i]$ for $i = 1, 2, 3, \ldots, n\}$ of configurations of the joint variables. The robot arm is an open- looped kinematic chain of links connected by prismatic or rotational joints. A local frame is fixed with each of the links and defines its relative position with respect to the previous link in terms structural parameters introduced by Denavit- Hartenberg notations. [17] [18]. This set of frames and structural parameters allows to define the forward kinematics of a robot arm mapping $F : Int\ Q_M \to W$, where W denotes the workspace of the robot arm. Assume the robot arm has redundant degrees of freedom ($n \geq m$) and it is capable of executing the preplanned work operation in task space. Without loss of generality we assume that $(\Lambda) \subset W$, where the workspace W of the robot arm is a subset of the task space R^m.

Definition. A path γ that satisfies $\gamma(\Lambda) \subset W$ we refer to as feasible path for the robot arm.

Denote the motion in the joints that allows the tool in the end- effector to execute the manipulation task subject to given (1) and (2) as follows

$$q(t) = (q_1(t), \ldots, q_n(t)): \mathrm{t} \in [t_0, t_1] \tag{4}$$

In case the path γ is a feasible path for the robot arm then there exists a robot arm motion (4) that executes the manipulation task subject to (1) and (2).

3 Separation of Concerns in Motion Path Planning

The object oriented approach for modeling is widely accepted among the software industry because it allows to introduce necessary levels of abstraction in dealing with change. Changes are inevitable and especially in an industrial environment, where production has to adapt to dynamically changing business requirements. For instance, with minimum efforts the robot arms must be reprogrammable to execute new tasks, or the tasks must be executed by a robot arm with a different structure. Hence, the software implementation of the model for motion control of a redundant robot arm along a given path in task space must be enough flexible to deal with variability among platforms and algorithms for motion control.

A typical problem in robotics is the difficulty in incorporating existing mathematical models [17] into a flexible software architecture built on reusable and evolvable software components. Recent research shows that this kind of architectures have become more and more relevant in the development of modern and future robotic systems [4] [19]. The inherent complexity of mathematical models in robotics is one of the primary reasons for transforming them into software applications with modern software architecture. An object oriented approach helps to decompose complex models on the basis of outlining independent areas of interest and creating a layered structure of subsystems. These independent areas of interest are known as concerns in software engineering.

The identification of concerns in software design is known as separation of concerns [20]. The Principle of separation of concerns is a fundamental principle in software engineering. It is used not just in software engineering rather appears in many different forms in the evolution of all methodologies, programming languages and best practices. The proposed solution in this paper makes use of Separation of concerns by separating the geometrical structure descriptions from the algorithms for motion control. This way it enables software applications of the model a software architecture that can provide support for various scenarios for adapting to changes. Moreover, this object- oriented approach allows to obtain new results in motion control, such as visualization of the Null space of a redundant robot arm and resolving Jacobian singularities.

The here considered problem is suitable to illustrate the application of the Separation of concerns principle in robotics. The function of motion $\lambda(t)$ (3) and the laws $q(t)$ (4) governing the motion in the joints represent the motion area of interest. We address this area of interest by representing $q(t)$ and $\lambda(t)$ in a common vector space. For uniformity in the presentation we denote $q_{n+1} = q_{n+1}(t) = \lambda(t)$.

Definition. The set $Q = Int\ Q_M \text{ x } R$ we call extended space of configurations.

Consider the motions $q^*(t) = \big(q(t), q_{n+1}(t)\big)$, $t \in [t_0, t_1]$ of the robot arm in the extended space Q that result in an end effector motion along the path γ. All such motions belong to the smooth manifold

$$\mathcal{B} = \{ q^* = (q, q_{n+1}) \colon F(q) - \gamma(q_{n+1}) = 0,\ q^* \in Q \} \tag{5}$$

where $q_{n+1} = q_{n+1}(t)$ represents the motion of the end effector along the path γ.

This way the Separation of concerns principle allows to represent the motion over the manifold $\mathcal{B}$ in the extended space of configurations in a similar way the motion occurs over the path γ in task space. We have two distinct areas of interest here. Every motion of the end- effector over γ has a corresponding motion executing over the manifold $\mathcal{B}$. Note that γ and $\mathcal{B}$ are independent areas of interest with respect to motions $\lambda(t) = q_{n+1}(t)$ and $q^*(t)$.

4 Visualization of Configurations in Null Space

The mathematical model for motion path control of a redundant robot arm is presented in a set of related papers [9] [21]. Here we will use this model to visualize the Null space of the robot arm. The Null space is treated often in research paper related to redundant robot arms, without success in visualizing it [22] [10]. The adopted object-oriented approach for investigating redundant robot arms enables us to understand visually the mechanism of redistributing the end effector motion internally among the joints of the redundant robot arm.

The mathematical model is a Jacobian- based model and makes use of vector space methods. It is derived from the equation for a vector $u^* = (u^T,\ u_{n+1})^T$:

$$J(q)u - {}^{\partial\gamma}\!/_{\partial q_{n+1}}(q_{n+1})u_{n+1} = 0 \tag{6}$$

where $J(q) = D(F)/D(q)$ is the Jacobian matrix of the forward kinematics $F(q)$. Any vector u^*that satisfies (6) belongs to the tangent vector space of $\mathcal{B}$ at point $q^* = (q, q_{n+1})$. Let J^+ be the Moore- Penrose pseudoinverse matrix of the Jacobian J and consider the following vector ξ:

$$\xi = J^+\ {}^{\partial\gamma}\!\Big/_{\partial(q_{n+1})} \tag{7}$$

For a fixed value of u_{n+1} the solutions u of (6) represent a linear manifold

$$L_{u_{n+1}} = \ \xi u_{n+1}\ +\ \aleph(J) \tag{8}$$

where $\aleph(J)$ is the linear subspace $\aleph(J) = \{u\ \in R^n{:}\,Ju = 0\}$ defined at any point of the smooth manifold. Let E denotes the unity $n\ x\ n$ matrix and define the following matrix:

$$P = E - J^+J \tag{9}$$

Matrix P is known as the projector matrix, because $Pu\ \in\ \aleph\big(J(q)\big)$. This matrix is symmetric $n\ x\ n$ matrix and it has the property

$$P^2 = P \tag{10}$$

Then for a given q_{n+1} we can use the following representation for the tangent vector field at $q^* = (q, q_{n+1})$ of $\mathcal{B}$:

$$L_{u_{n+1}} = \ \xi u_{n+1}\ + Pu \tag{11}$$

On the other hand for a fixed $q_{n+1} = \bar{\lambda}$ the Null space of the robot arm can be represented on Figure 2 as:

$$\mathcal{B}_{q_{n+1}} = \{q\ \in Int\ Q_M{:}\,F(q) - \gamma(q_{n+1}) = 0\} \tag{12}$$

Here $\mathcal{B}_{q_{n+1}}$ is the intersection of $\mathcal{B}$ with the plane $q_{n+1} = \bar{\lambda}$ and consists of all the configurations q of the robot arm for which the end effector position retains the position $\gamma(\bar{\lambda})$, $\bar{\lambda} \in [\lambda_1, \lambda_2]$.

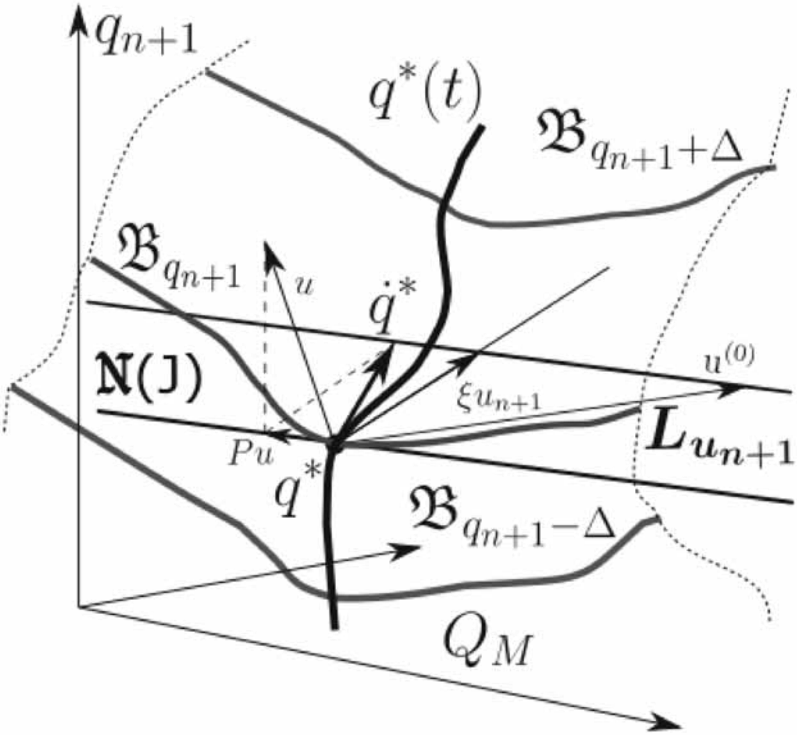


Fig. 2. Visualization of the Null space

Figure 2 displays the trajectory of motion $q^*(t)$. It lies entirely on $\mathcal{B}$. Clearly, vector $\dot{q}^*$ belongs to the tangent vector field of $\mathcal{B}$ at q^* and it is represented in $L_{u_{n+1}}$by vectors Pu and ξu_{n+1}. This figure has the following important interpretations for the roles vectors Pu and ξu_{n+1} in executing a redundant robot arm motion.
Vector ξu_{n+1} is responsible for moving the end- effector of the robot arm from one point to another point on path γ.

1. Provided Pu is not the zero vector, while at the same time $u_{n+1} = 0$, then the robot arm motion will result in reconfiguring the robot arm without changing the end- effector position in task space. The velocity of the end- effector is determined entirely by the value of u_{n+1}.
2. The velocity of the end- effector depends only on the value of u_{n+1}. The larger u_{n+1} is, the larger is the vector of the end- effector velocity
3. In case $u_{n+1} < 0$ then the end- effector motion proceed toward the starting point of its path γ in task space.
4. Vector Pu causes the robot arm to redistribute internally the robot arm motion by adding a component to ξu_{n+1} that results in Null space motion. This

way, part of the end- effector motion is “wasted” on reconfiguration. Note that the component Pu in Null space will not change the velocity of the end-effector in terms of size and direction.

Figure 2 provides valuable insight about how the joint motion of a redundant arm proceeds. At any given moment of motion in task space it shows the motion in the joints in relation to the current Null space. We make use of this representation to develop a heuristic procedure processing Jacobian singularities during motion path control.

5 Passing- through Singularities

Jacobian singularities are a subject of active research for many years. The solution of the inverse kinematic problem relies on the assumption that the rank of the Jacobian is equal to the dimension of the task space for all the configurations of joint variables. In practice, there exist singularities q, where $rank\ J(q) < m$. These singularities emerge as a major obstacle in finding a general solution for the inverse kinematic problem in robotics by means of a Jacobian- based models. Singularities cause a discontinuity in the joint space solution for the inverse kinematic problem. The rank of the Jacobian falls in a singularity and as a result decreases the number of the independent direction for displacement of the end- effector in task space. Therefore a robot arm is difficult to control at a singularity [12].

The research in this direction can be summarized as classifying singularities, developing criteria for identifying and avoiding singularities. Most often the classification of singularities refers to particular robot arm structures like PUMA 560 and relies on the symbolic representation of the forward kinematics [23]. The manipulability index is employed as a measure for the distance to singularities for the purpose of avoiding them [24]. At the same time it is established that not all singularities can be avoided especially those of orientation origin [18]. To avoid the inversion of the Jacobian series of task resolution techniques have been proposed [25]. Singular value decomposition and damped least- squares are other frequently used techniques for handling singularities [13] [14] [26].

Instead of avoiding the here proposed method allows to pass- through a Jacobian singularity in the context of motion control of redundant robot arms. This method exploits substantially both the Separation on concerns principle [27] and the Continuity principle [28]. Unlike other approaches to the solution of this problem we use substantially the assumption that the end- effector motion executes over a fixed geometric path in task space and it represents an independent concern with respect to this path. We consider similarly the robot arm motion in the extended space of configurations, where the motion executes also over a geometric object, the manifold $\mathcal{B}$. Without loss of generality we also assume that the path γ is a feasible path for the robot arm. Note that the currently existing solutions make no separation between these two major areas of concern.

The problem of resolving Jacobian singularities relates to the problem of identifying such singularities. Consider the projector matrix P introduced in (9). Its $rank\ P(q) = n - m$ for all , where $rank\ J(q) = m$. Moreover, the projector matrix P is a symmetric $n \times n$ matrix for which the property $P^2 = P$ (10) holds true. The trace of a matrix is invariant with respect to the change of basis. Now from the Spectral theorem in algebra [29] it follows that P has exactly $n-m$ orthonormal eigenvectors. On the other hand, the trace of P is the sum of its eigenvalues. The sum of their corresponding eigenvalues is equal to $n-m$. This proves the following proposition.

Proposition 1. The Jacobian has full rank, if and only if, the trace of matrix P is equal to $n-m$.

This proposition provides an efficient numerical criterion to identify a singularity of the Jacobian. It eliminates the need to estimate the proximity to a singularity or the use of singular value decomposition.

Corollary 1. A configuration q is a Jacobian singularity, if and only if, the trace of matrix P is larger than $n-m$.

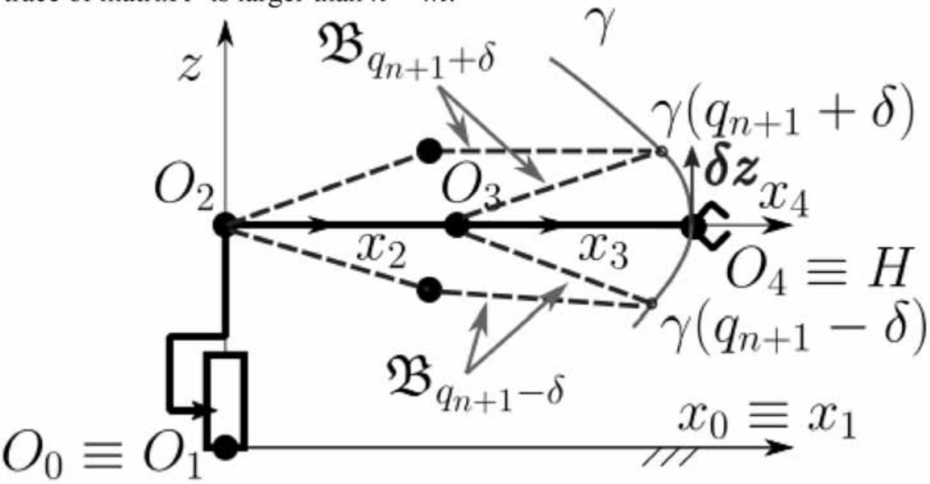


Fig. 3. Singular configuration of a robot arm.

Once we have established a criterion to identify a singularity, let us consider a Jacobian singularity on Figure 3. Select a small neighbourhood $\delta > 0$ around the singular configuration. Denote by $\gamma(q_{n+1})$ the position of the end-effector in work space W such that the configuration of the robot arm O_1, O_2, O_3, O_4 is a Jacobian singularity. The mobility of the end- effector is decreased in case of a singularity. For instance, in this example the end- effector mobility is reduced from $\delta x, \delta z$ to δz. Note, the end- effector can continue its motion on the assumption γ feasible path for the robot arm because the tangent vector of γ is collinear with the mobility direction . Hence, there exists a robot arm motion $q^*(t) = (q(t), q_{n+1}(t))$ that can pass through the singularity configuration.

Consider now an admissible motion $q^*(t)$ of the robot arm in the extended space of configurations that corresponds to the end- effector motion on γ in task space (Fig. 4). For clarity, denote by $q^{(s)}$ the Jacobian singularity at $\gamma(q_{n+1})$.

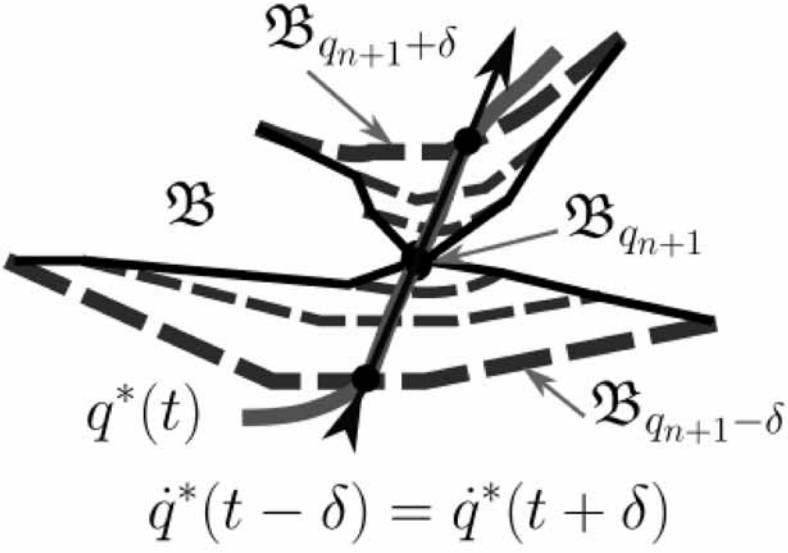
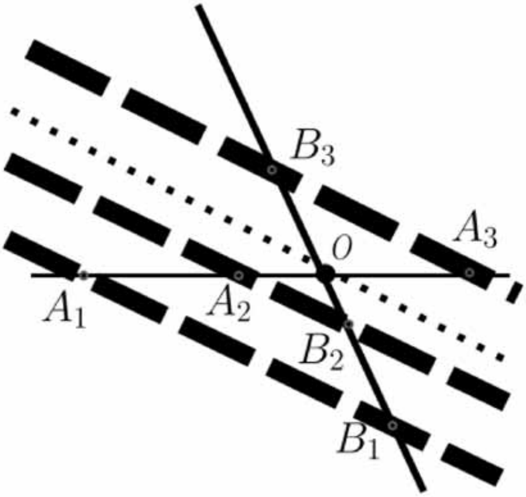

$\mathfrak{B}_{q_{n+1}+\delta}$
$\mathfrak{B}$
$\mathfrak{B}_{q_{n+1}}$
$q^*(t)$
$\mathfrak{B}_{q_{n+1}-\delta}$
$\dot{q}^*(t-\delta) = \dot{q}^*(t+\delta)$

Fig. 4. Motion in extended space of configurations passing through singularity.

This motion executes on $\mathcal{B}$ and the singularity configuration $(q^{(s)}, q_{n+1})$ also belongs to $\mathcal{B}$. Note that $q^*(t)$ passes through Null spaces $\mathcal{B}_{q_{n+1-\delta}}$ of dimension $n-m$ before it reaches the singular configuration. The tangent vector field on $\mathcal{B}$ at $q^*(t) = (q^{(s)}, q_{n+1})$ becomes undefined this singular configuration. From Corollary 1 it follows that $rank\, P\big(q^{(s)}\big) > n-m$ in the singular configuration. In fact the decrease in the mobility of the end- effector leads to an increase in the dimension of the Null space in a singular configuration. This increase in the dimension of the Null space has no effect on the mobility of the end- effector in task space. As noted in Section 4, vectors from the Null space cannot change the direction and the size of the velocity vector of the end- effector. Therefore this change in the dimension of the Null space in a singular configuration won't affect the motion of the end effector.

In the general case, the singularity $q^{(s)}$ separates one set of non-singular configurations from another set of non-singular configurations. This situation is similar to the example used by Leibnitz to formulate the Principle of continuity [28]. In this example the fraction of the segments$A_iO, B_iO, i = 1, 2, 3$ is defined by continuity in point O as equal to ${}^{A_iO}/_{B_iO}$. (Fig. 5).



$$A_1O/B_1O = A_2O/B_2O = OO/OO$$

Fig. 5. Example illustrating the Principle of Continuity.

In terms of the Principle of continuity we can interpret the singularity as a "gate" between sets of non- singular configurations on $\mathcal{B}_{q_{n+1}}$. Hence, we can replace the tangent vector field at $(q^{(s)}, q_{n+1})$ in $\mathcal{B}$ with the tangent vector field at point (q, q_{n+1}) in $\mathcal{B}$ selected in a sufficiently small neighbourhood of $(q^{(s)}, q_{n+1})$ such that $rank\, J(q) = m$. In order to pass through this "gate" we define by continuity the velocity of the joint motion in the singular configuration to be equal to its velocity for a non- singular configuration selected from a close neighbourhood of the singular configuration. Indeed, a small variation in the configuration $\delta q^{*(s)} = (\delta q^{(s)}, \delta q_{n+1})$ in the extended space of configurations would cause a related displacement $\delta\gamma$ of the end- effector in task space corresponding to the variation δq_{n+1}. On the assumption that γ is a feasible path for the robot arm then the resulting $\delta\gamma$ displacement will belong to the space of mobility directions.

These findings allow us to conclude that the robot arm motion can "passing through" singularities in motion control on the assumption that the geometric path γ is feasible for the robot arm. This assumption for feasibility of the path γ is a natural requirement for a robot arm to execute a work task. The obtained results correspond to the physical observations about how a robot arm passes through a singularity. In fact the Jacobian singularity is a pure mathematical problem. It is not a problem that blocks the robot arm motion and the observations in practice prove it. Finally, we wouldn't be able to obtain these results without the application of the Separation of concerns principle.

6 Conclusion

This paper focuses on the application of an object- oriented approach for motion control of redundant robot arms along a path in task space. We demonstrate that blending methods in software engineering with mathematical methods can produce new solutions to existing problems in robotics. In the beginning we the problem of task planning from an object oriented viewpoint. Further on we apply the Separation of concerns principle to interpret the major types of concerns in an extended space of configurations of a redundant robot arm. This allows us to treat the robot arm motion in the extended space of configurations in a similar way the end- effector motion executes in task space. Further on, the selected approach enables a visual representation of the Null space of a redundant robot arm in the context of its motion. The visual interpretation of the Null space is used to solve the Jacobian singularity problem. This problem is of major concern for kinematic control of manipulators. An efficient numerical criterion is proposed to identify a Jacobian singularity. The Jacobian singularity problem is solved in terms of separating areas of concerns and applying the Principle of continuity.

References

1. A. Angerer, A. Hoffmann, F. Ortmeier, M. istein and W. Reif, "Object-Centric Programming:A New Modeling Paradigm for Robotic Applications," in *ICAL'09. IEEE International Conference on Automation and Logistics*, 2009.
2. A. Angerer, A. Hoffmann, A. Schierl and M. Vistein, "Robotics API: Object-oriented software development for industrial robots," *Journal of Software Engineering for Robotics,* vol. 4, no. 1, pp. 1-22, 2013.
3. A. Angerer, Object-oriented Software for Industrial Robots, Augsburg, Germany: Doctoral dissertation, University of, 2014.
4. I. Nesnas, R. Simmons, D. Gaines, C. Kunz, A. Diaz-Calderon, T. Estlin, R. Madison, J. Guineau, M. McHenry, I. Shu and D. Apfelbaum, "CLARAty: Challenges and steps toward reusable robotic software," *International Journal of Advanced Robotic Systems,* vol. 3, no. 1, pp. 23- 30, 2006.
5. E. Prassler, H. Bruyninckx, K. Nilsson and A. Shakhimardanov, "The use of reuse for designing and manufacturing robots.White Paper," Robot Standards and reference architectures (RoSta) consortium, 2009.
6. I. Patias and V. Georgiev, "Traffic Prioritization System Based on Embedded Components," in *9th International Conference ISGT'2015 (Information Systems & Grid Technologies)*, Sofia, Bulgaria, 2015.
7. I. Patias and V. Georgiev, "Embedded Architecture of Tolls Collecting System," in *9th International Conference ISGT'2015 (Information Systems & Grid Technologies)*, Sofia, Bulgaria, 2015.
8. B. Bruegge and A. H. Dutoit, Object-Oriented Software Engineering Using UML, Patterns, and Java, Prentice Hall, 2010.
9. E. Krustev and L. Lilov, "Kinematic path Control of Robot Arms with Redundancy," *Technische Mechanik,* vol. 6, no. 2, pp. 35- 42, 1985.
10. F. Flacco, A. De Luca and O. Khatib, "Motion Control of Redundant Robots under Joint

Constraints: Saturation in the Null Space," in *IEEE International Conference on Robotics and Automation (ICRA)*, St. Paul, MN, USA, 2012.

11. H. C. Cho and J. B. Song, "Null space motion control of a redundant robot arm using matrix augmentation and saturation method," in *12th International Conference on Motion and Vibration Control*, Sapporo, Hokkaido, Japan, , 3-7 August, 2014.
12. T. Shamir, "The Singularities of Redundant Robot Arms," *The International Journal of Robotics Research,* vol. 9, no. 1, pp. 113-121, 1990.
13. S. Chiaverini, G. Oriolo and I. D. Walker, "Kinematically Redundant Manipulators," in *Handbook of Robotics*, B. Siciliano and O. Khatib, , pp. 245- 265, Springer, 2008.
14. B. Tondu, "A Theorem on the Manipulability of Redundant Serial Kinematic Chains," *Engineering Letters,* vol. 15, no. 2, p. 362, 2007.
15. B. Siciliano, L. Sciavicco, ,. L. Villani and G. Oriolo, Robotics: Modelling, Planning and Control, 1 ed., Springer Publishing Company, Incorporated, 2008.
16. G. Calafiore and L. E. Ghaoui, Optimization Models, Cambridge University Press, 2014.
17. L. Lilov and G. Bojaddziev, Dynamics and Control of Manipulative Robots, Sofia: St. Kliment Ohridski University Press, 1997.
18. F. C. Park and K. M. Lynch, Introduction to Robotics. Mechanics, Planning and Control, Cambridge University Press, 2017.
19. I. Stanev, "Method for Automated Programming of Robots," in *Knowledge Based Automated Software Engineering*, Cambridge, Cambridge Scholars Press, p. 67 – 85,2012,.
20. ISO/IEC, Software Engineering – Guide to the Software Engineering Body(SWEBOK). Version 3, P. Bourque and R. E. (. Fairley, Eds., IEEE, 2014.
21. E. Krastev, "Mathematical Model for Motion Control of Redundant Robot Arms," *WSEAS Transactions on Systems,* vol. 16, pp. 36-42, 2017.
22. G. Antonelli, F. Arrichiello and S. Chiaverini, "The Null-Space-Based Behavioral Control for Autonomous Robotic Systems," *Intelligent Service Robotics,* 1(1), pp. 27- 39, 2008.
23. M. J. Mirza, M. W. Tahir and N. Anjum, "On Singularities Computation and Classification of Redundant Robots," in *International Conference on Computing, Communication and Control Engineering (IC4E'2015)*, Dubai (UAE), 2015.
24. T. Yoshikawa, "Manipulability of robotic mechanism," *The International Journal of Robotics Research,* vol. 4, no. 2, p. 3–9, 1985.
25. S. Chiaverini, "Singularity-Robust Task-Priority Redundancy Resolution for Real-Time Kinematic Control of Robot Manipulators," *IEEE Transactions on Robotics and Automation,* vol. 13, no. 3, pp. 398- 410, 1997.
26. S. R. Buss, "Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods," *IEEE Journal of Robotics and Automation* , vol. 17, pp. 1-19 , 2004.
27. E. W. Dijkstra, A Discipline of programming, Englewood: Prentice Hall, 1976.
28. G. W. Leibniz and L. E. Loemker, Philosophical Papers and Letters, 2 ed., R. Ariew and D. Garber, Eds., Indianopolis & Cambridge: Hackett Publishing Company Dordrecht: D. Reidel, 1989, p. 546.
29. J. Solomon, Numerical Algorithms: Methods for Computer Vision, Machine Learning, and Graphics, CRC Press , 2015.

Formal Specification of CAPECs in UML & OCL

Vladimir Dimitrov,

Faculty of Mathematics and Informatics, University of Sofia, 5 James Bourchier Blvd.,
1164 Sofia, Bulgaria
cht@fmi.uni-sofia.bg

Abstract. Cyber-attacks are described in formatted text. There is no widely accepted formal notation for that purpose. This paper shows how UML can be used for formal specification of CAPEC-47.

Keywords: cyber-attack, formalization, UML.

1 Introduction

Cyber-attacks are described in CAPEC database [4]. The cyber-attack exploits weakness or weaknesses to achieve a vulnerability. First, the vulnerability is detected, then after investigation and analyzes the attack that has been performed is detected and the corresponding weaknesses that are exploited are discovered. So, weaknesses are classified in CWE [2] and vulnerabilities in CVE [1].

An example of such attack is CAPEC-47 given below:

“

CAPEC-47: Buffer Overflow via Parameter Expansion

Attack Pattern ID: 47

Abstraction: Detailed

Status: Draft

Completeness: Complete

Presentation Filter: Basic Complete

Summary

In this attack, the target software is given input that the attacker knows will be modified and expanded in size during processing. This attack relies on the target software failing to anticipate that the expanded data may exceed some internal limit, thereby creating a buffer overflow.

Attack Prerequisites

- The program expands one of the parameters passed to a function with input controlled by the user, but a later function making use of the expanded parameter erroneously considers the original, not the expanded size of the parameter.
- The expanded parameter is used in the context where buffer overflow may become possible due to the incorrect understanding of the parameter

size (i.e. thinking that it is smaller than it really is).

Solutions and Mitigations

Ensure that when parameter expansion happens in the code that the assumptions used to determine the resulting size of the parameter are accurate and that the new size of the parameter is visible to the whole system

Related Attack Patterns

Nature	**Type**	**ID**	**Name**
ChildOf	S	100	Overflow Buffers

“

There are several approaches to describe attack patterns discussed in another papers. Here the focus is the specification of cyber-attacks with UML [4].

2 CAPEC-47 in UML

This attack is so called “Heartbleed” and in the specification is included some knowledge from CVE-2014-0160 and CWE-119 that describe the attack from vulnerability and weakness point of view.

First, the two signals for a Heartbleed are defined in the class diagram below in Fig. 1.

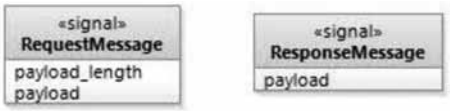


Fig. 1. SIngals.

The attribute types are not defined, because nowhere in CWEs or CVEs and CAPECs is mentioned anything about their types. They have to be C types as follows, independently of attacker software:

```
payload_length: short;
payload: string;
```

In addition for the RequestMessage to be an Attacker message, an OCL constraint must be put on the RequestMessage signal such that the payload length must be less than payload_length – that the attack essence.

The attack is specified with an activity diagram as follows in Fig. 2.

The attacker create an appropriate message and send it to the target software.

Here CWE-119 is a software weakness CWE-119 that is exploited by the attack. The last one receives the message and expands the parameter using payload_length but not the actual size of the payload. This means that the received payload is loaded in a buffer that is bigger than the actual size of the payload. Therefore after the received payload there are some uninitialized content that can contains sensible information.

The target returns expanded parameter to the attacker using the payload_length i.e. the full buffer containing the received payload and the uninitialized part. The attacker receives the response message and process the sensible information.

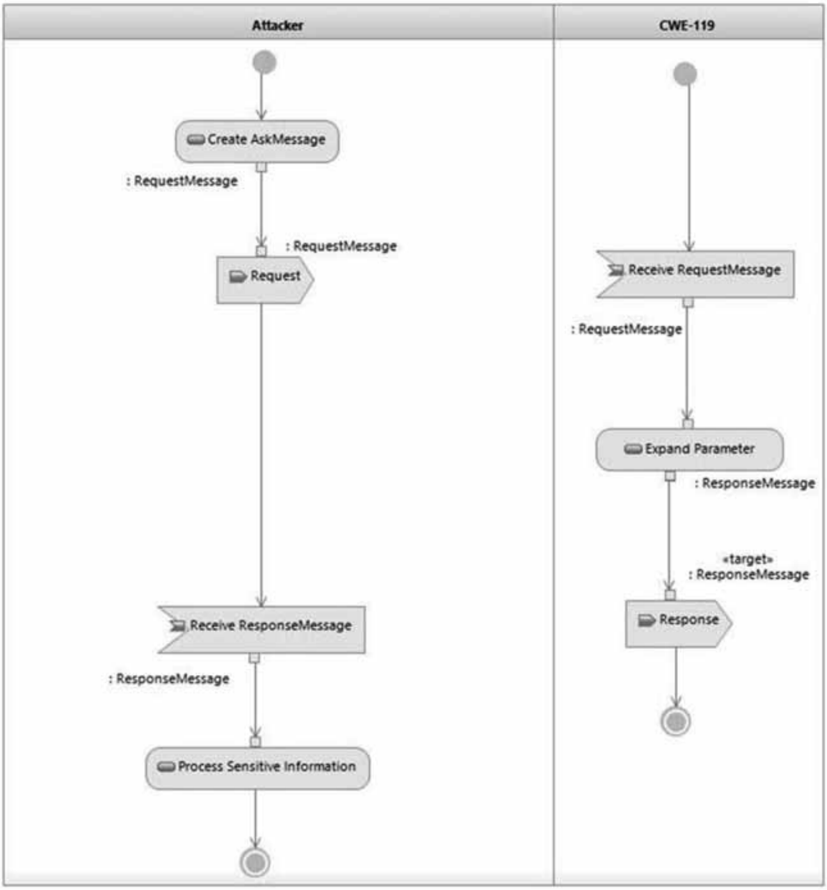


Fig. 2. CAPEC-47.

Another more detailed version of this attack is given in Fig. 3.

Here an activity is included in the target explaining the weakness CWE-119 essence.

The Attacker process and the Server with CWE-119 process are running independently. The Attacker creates its RequestMessage and sends it to the CWE-119 process. The server process receives the Attacker message via the Receive signal event that triggers its execution. The action "send RequestMessage" and the trigger "receive RequestMessage" are hidden in the corresponding activities as properties.

The server expands the Payload parameter, creates Response message, and answer back to the Attacker.

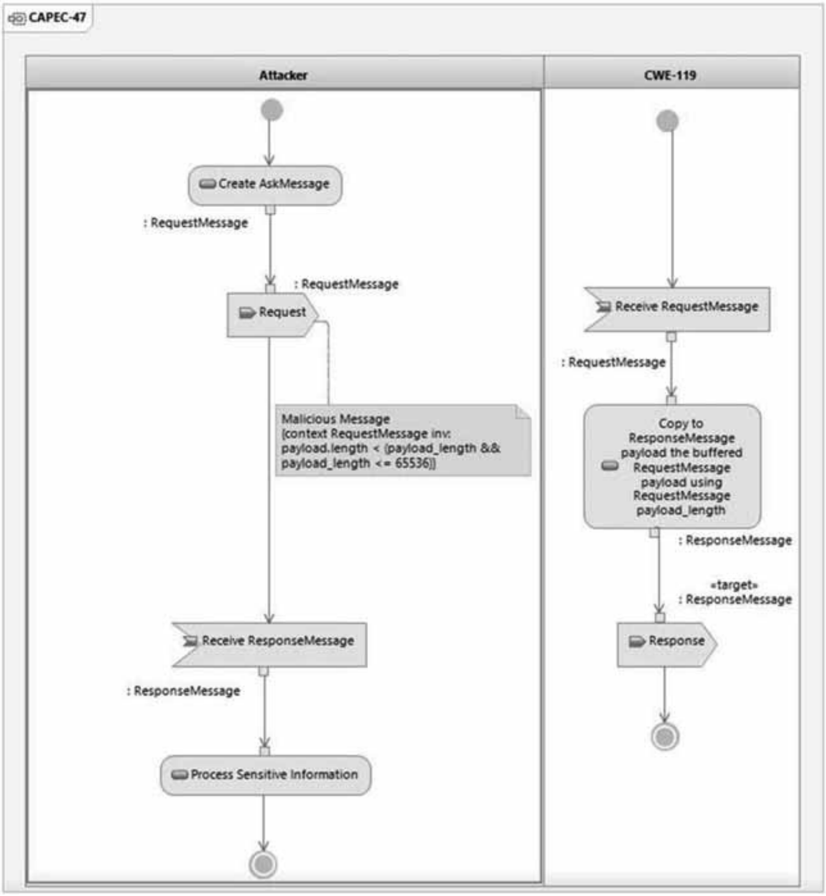


Fig. 3. CAPEC-47 more detailed version.

In action "Expand parameter" is added OCL constraint that the payload of ResponseMessage has RequestMessage payload_length and that it contains the payload:

```
   context RequestMessage inv: payload.length <
(payload_length && payload_length <= 65536)
```

3 Conclusion

Furthermore, the project can be converted to C++ project and code generated.

In IBM Rational Software Architect is possible a UML model project to be imported in a development project (C++ project, Java project etc). The last one can be used for test code generation. The import preserves all source model constraints including OCL ones and they are used in the following code generation.

OCL specification can be used to connect the CWE-119 Z-specification to the UML class diagram and the UML activity diagrams. The problem here is how to convert a Z-specification to OCL one. What exactly have to be converted and how this must be done? Theoretically it is possible.

The chains of CWEs, can be specified as CAPECs in UML.

References

1. MITRE. “CVE Common Vulnerabilities and Exposure.” http://cve.mitre.org
2. MITRE. “CWE Common Weakness Enumeration.” http://cwe.mitre.org
3. MITRE. “Common Attack Pattern Enumeration and Classification”, http://capec.mitre.org
4. OMG, UML, http://www.uml.org

Bulgarian e-Government Information System Based on the Common Platform for Automated Programming – Requirements

Ivan Stanev, Maria Koleva

Faculty of Mathematics and Informatics, University of Sofia St. Kliment Ohridski
5 James Bourchier blvd., 1164 Sofia, Bulgaria
instanev@fmi.uni-sofia.bg, mkoleva@fmi.uni-sofia.bg

Abstract. Challenges to the Bulgarian e-Government information system (BeG) related to development process, semantic interoperability, and authorization management are described. Leading e-government solutions including domain independent and specialised components are analysed and compared. Important BeG requirements are specified related to automation, scalability, flexibility, reusability, etc. BeG concept is developed. The main principles for BeG realisation are identified.

Keywords: SOA, Cloud computing, Knowledge based automated software engineering, common platform for automated programming, e-government.

1 Introduction

The European Commission programming periods 2007 - 2013 and 2014 - 2020 are performed with a focus on e-governance introduction in our everyday life. Developments are so intense that stakeholder requirements usually exceed the capabilities of the software development tools used.

Hereafter are presented the important problems, the selected solutions, the results achieved, and the difficulties faced by the management team[1] responsible for e-government in Bulgaria (BeG) in the period 2010 - 2013 and the e-government development planning for the period 2014 – 2020.

At the beginning of this period, the state of BeG detailed in [CoM, 2016] and [NIFO, 2016] can be represented by numbers as follows:

(1) BeG Portal provides access to 1300 services which offer (often outdated) information on service formalities; (2) 0 services of interaction and transaction type; (3) 27% of all 568 registers are maintained only on paper; (4) 8% of all state

[1] The authors of the article were part of the e-government management team in the period 2010 - 2013.

V. Dimitrov, V. Georgiev (Editors): ISGT'2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

registers exchange data with partner organisations; (**5**) 12% of all administrations are registered to exchange electronic documents through the Electronic data exchange environment. No document is exchanged in this environment; (**6**) 44% of all 539 administrations (central, local) do not have a database management system.

These data show that the development of e-government in Bulgaria practically started at that time.

2 Bulgarian e-Government Goals

The main objective of BeG development is illustrated by the description of an electronic administrative service in two versions – As Is - as implemented in the beginning of the period and To Be - as it should be implemented at the end of this period.

Either citizen or business organisation, we interact with the public administration on a regular basis – for registrations, payments, life events, etc. Our expectations are that it would be possible to do it by electronic means, thus saving our own as well as government resources.

Electronic services delivery is governed by the E-government Act (http://lex.bg/laws/ldoc/2135555445) adopted in 2007. One of the Act's main provisions is that administrative bodies cannot require from citizens and organisations to produce, or to prove data which has already been collected or created. Such data must be collected ex officio.

One of the priority groups of services for businesses identified by the European Commission and Member States under the Your Europe Initiative (http://europa.eu/youreurope/) is VAT and Customs. The National Customs Agency is one of the largest providers of electronic services in Bulgaria [CoM, 2016].

We present hereafter a case study for the licensing of a tax warehouse where excise goods are processed under duty suspension arrangements.

According to the Excise Duties and Tax Warehouses Act (http://www.lex.bg/laws/ldoc/2135512728) the economic operator shall apply for a license authorizing him to manage a tax warehouse. His application shall be complemented with a number of documents, i.a. registration for business operations, conviction status certificate, tax liabilities certificate, sketches, etc.

The economic operator is required to obtain these documents from the Registry Agency, the National Revenue Agency, the Geodesy, Cartography and Cadastre Agency, etc. prior to applying for his tax warehouse license. This process is presented in Fig. 1

In Fig 2. is presented the solution which must be operational at the end of the period. This solution is characterized by a one-stop shop service, whenever possible in real time. In that case the economic operator applies for a license through electronic means. The application is processed by a customs official, all the necessary checks are performed via G2G links with other public organisations, and the license is prepared and submitted electronically to the economic operator. Other channels of communication shall be kept open for those who are disconnected by choice or necessity.

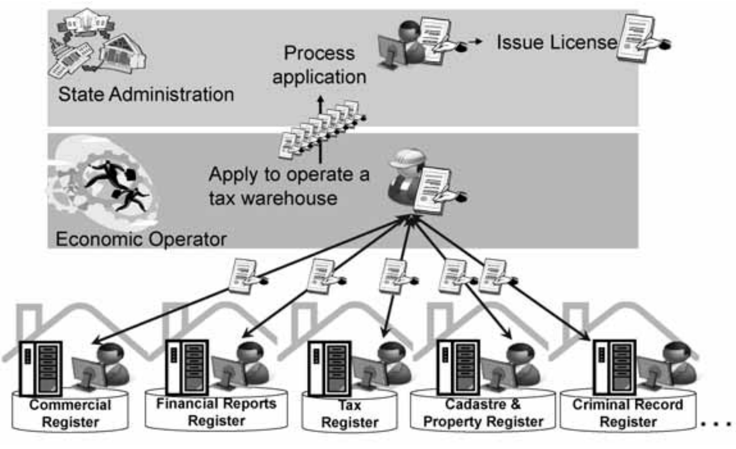


Fig. 1 Issue License for Tax Warehouse Process – Case Study As Is

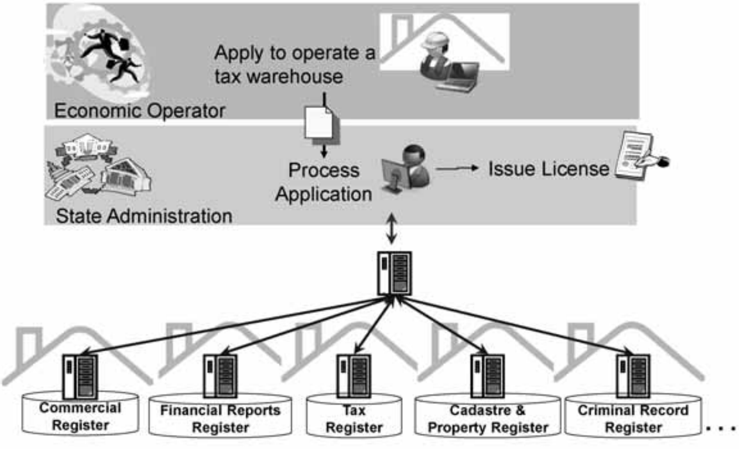


Fig. 2 Issue License for Tax Warehouse Process – Case Study To Be

There are significant differences between the As Is and To Be processes. In the As Is process are involved at least 6 employees, the economic operator has to make at least 7 visits to various institutions, there are long waiting periods between the different steps in the process. In the To Be process 1 employee is engaged, no site visit is needed, there is virtually no waiting time between the different steps in the process.

As a result, general BeG Goals can be identified as follows: (**1**) reduction of user resources – save time, and money of the business and citizen; (**2**) improvement of state administration efficiency – optimization of the operation time, and required resources (financial, human, equipment, etc.) for the state administration activities; (**3**) more security – reduction of the risks of fraud, through increased automatic validations and verifications.

In order to achieve these goals the following **important problems**, identified in the presented Case study, shall be overcome: (**1**) lack of single access point with a centralized authorization solution; (**2**) big volume of not digitized data and low quality of digitized data; (**3**) low level of semantic interoperability, including lack of knowledge for determined data interpretation; (**4**) lack of instruments for fast information systems adaptation to rapidly changing legal base.

These problems form the basis for the comparative criteria used in the state of the art review hereafter.

3 E-Government State of the Art

The review done covers platforms realised by some of the leaders in e-government development among the member states of the United Nations [UN, 2016]. Seven e-government platforms (see Table 1) are selected, compared and analysed: (**1**) the Australian Government Architecture Framework Reference Models [AGIMO 2011], [ADFD 2013], (**2**) the National Architecture for Digital Services in Finland [FPRC, 2016], (**3**) the Swedish e-government system [NIFO, 2016], (**4**) the Dutch Government Reference Architecture [Dutch Government, 2010], (**5**) the New Zealand Government Enterprise Architecture Reference Models and Taxonomies [New Zealand Government, 2015], (**6**) the Estonian e-government system [Vassil, 2015], https://e-estonia.com/components/, https://www.ria.ee/en/], and (**7**) the Austrian e-government system [Austrian Federal Chancellery, 2014].

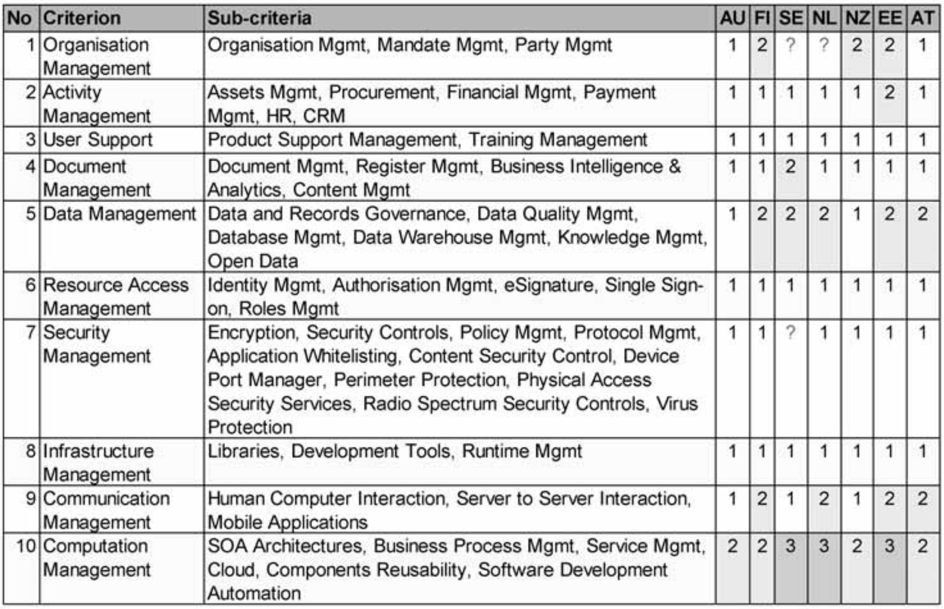
These platforms have been selected due to similarities with the Bulgarian e-government legal framework (e.g. Austria), for the diversity in systems development approach (e.g. Australia, New Zealand, Finland, Denmark), or because of comparable problems faced (e.g. Estonia).

The results of the comparative analysis are shown in Table 1. In the rows of the table are presented government business and SOA components. In columns with two-letter country code are included the compared e-government frameworks (AU – Australia, FI – Finland, SE – Sweden, NL – Netherlands, NZ – New

Zealand, EE – Estonia, AT - Austria). At the intersection of rows and columns is presented one of the following symbols: 1 – “the framework has the corresponding characteristics”, 2 – “the framework has most of the characteristics in the sub-criteria”, 3 – “the framework has some of the characteristics in the sub-criteria”, ? – “no information is found on whether the framework has the corresponding characteristics.”

This analysis shows that: (**1**) all systems are developed with a focus on key business and infrastructure components, e.g. development and runtime environments (**2**) a clear trend is the realisation of the e-government technology enablers [Tinholt, 2015] such as e-identification, documents and registers management; (**3**) organisation management, data management and communication management functions are limited; (**4**) there is considerable shortage in computation management instruments, e.g. SOA, components reusability, software development automation.

Table 1 E-Government frameworks comparative analysis results



No	Criterion	Sub-criteria	AU	FI	SE	NL	NZ	EE	AT
1	Organisation Management	Organisation Mgmt, Mandate Mgmt, Party Mgmt	1	2	?	?	2	2	1
2	Activity Management	Assets Mgmt, Procurement, Financial Mgmt, Payment Mgmt, HR, CRM	1	1	1	1	1	2	1
3	User Support	Product Support Management, Training Management	1	1	1	1	1	1	1
4	Document Management	Document Mgmt, Register Mgmt, Business Intelligence & Analytics, Content Mgmt	1	1	2	1	1	1	1
5	Data Management	Data and Records Governance, Data Quality Mgmt, Database Mgmt, Data Warehouse Mgmt, Knowledge Mgmt, Open Data	1	2	2	2	1	2	2
6	Resource Access Management	Identity Mgmt, Authorisation Mgmt, eSignature, Single Sign-on, Roles Mgmt	1	1	1	1	1	1	1
7	Security Management	Encryption, Security Controls, Policy Mgmt, Protocol Mgmt, Application Whitelisting, Content Security Control, Device Port Manager, Perimeter Protection, Physical Access Security Services, Radio Spectrum Security Controls, Virus Protection	1	1	?	1	1	1	1
8	Infrastructure Management	Libraries, Development Tools, Runtime Mgmt	1	1	1	1	1	1	1
9	Communication Management	Human Computer Interaction, Server to Server Interaction, Mobile Applications	1	2	1	2	1	2	2
10	Computation Management	SOA Architectures, Business Process Mgmt, Service Mgmt, Cloud, Components Reusability, Software Development Automation	2	2	3	3	2	3	2

Important **conclusions based on comparative analysis** are: (**1**) all systems are developed with a focus on key business and infrastructure components, e.g. development and runtime environments; (**2**) realised are the e-government technology enablers such as e-identification, documents management and registers management; (**3**) the organisation management, data management and communication management functions are limited; (**4**) computation management instruments, e.g. SOA, components reusability, software development automation, are scarcely used.

4 BeG Requirements

Based on the State of the Art conclusions is defined a set of **important requirements** concerning the realization of BeG, including: (**1**) Digitisation of all important data; (**2**) high level of data quality control; (**3**) Standardised mechanism for automated exchange of information in computer-readable format; (**4**) Standardised interfaces for automated official access to registers; (**5**) Standardised modules for integration of sectorial IS with State registers; (**6**) Standardised BeG ontology; (**7**) Synchronised business processes – e.g. how to act once organisations have exchanged information; (**8**) Single access point with central authorization; (**9**) Support of Resources Register; (**10**) Support of BeG Officials Register; (**11**) Repository with reusable services and instruments, SOA; (**12**) development of Public Cloud.

5 Concept for Realisation

The Concept for BeG realisation is presented in three main groups of principles: general, organsiational, and technical.

The **General Principles** group includes: (**1**) **Complexity** – one service, or one business process with one I/O point. (**2**) **Automated information exchange** – data is collected or created once, and is reused many times. (**3**) **Unification and modularisation of e-services** – services and other components are developed once, and are reused many times.

The **Organisational Principles** group includes:

(**1**) **E-Government Registers Catalogue**- The e-government registers (eGR) and the registers of the primary administrator of the data (PA) are not the same. eGR represent an excerpt from the PA registers for which eGR users have legally justified access right. eGR are stored in either PA or BeG data center.

(**2**) **Integral interpretation of the Legal Base** - BeG shall be developed in accordance with the E-Government Act (EGA) and the sectoral legal base aligned with the EGA.

(**3**) **Best Practices studied** - BeG services shall be developed following sectoral and e-government best practices at national and international level.

(**4**) **Consensus based management** - eGR development shall be based on decisions taken by consensus by a joint team composed of representatives of the PA, the central e-governance unit (as of 01.07.2016 the competences of the E-Governance Directorate in the Ministry of Transport, Information Technologies and Communications have been transferred to the E-governance State agency, EGSA), and the supplier.

(**5**) **Primary administrator is master of the eGR development process** - eGR commissioning shall be authorized by EGSA following approval by the PA.
(**6**) **Common Project Management Approach** - A unified development process, such as the Rational Unified Process [IBM], shall be used for business modelling, requirements, analysis, design, implementation, testing, deployment, production and management of BeG software products.
(**7**) **Project preparation and implementation transparency** – tender dossiers shall be subject to public discussion, representatives of citizens and ICT organizations shall take part in deliverables acceptance.
The **Technical Principles** group includes: (**1**) **IT Systems Code Ownership** - the supplier must provide the source code of the BeG software product to the sectorial contracting authority. It is delivered to the EGSA as well, if such provisions are in place.
(2) New eService compatibility with the Infrastructure – eGR shall be developed to efficiently operate on the existing PA or EGSA infrastructure (as agreed between them). The PA shall be responsible for the quality of the software product. The owner of the infrastructure shall be responsible for data security.
(**3**) **Flexibility of the Developed eGRs** – the software realisation shall enable rapid and inexpensive modification of the relevant services if legal, organizational or technical changes occur.
(**4**) **System Design before Law Construction** – eGR shall be developed so as to achieve the best possible efficiency. If necessary, proposals for modifications in the existing legal base shall be prepared. Proposed modifications shall be incorporated in system specifications, models, and realisation. eGR shall be deployed for production after the entry into force of the modified legal base.
(**5**) **eGR Standardisation** - The supplier shall prepare eGR for registration in accordance with the EGA, prepare standard protocols for data exchange, develop tests suites and test procedures. New users of the eGR are joined following an open process described in the prepared documentation and published for all interested organisations.
The **Software Architecture of the Concept** presented in Fig.3 includes: (**1**) **Zone 1** **End Users** - End users of the system, getting access to individual data after identification and authorization. These users are provided with a personal working space in the BeG Portal to draft and store the required information, if they need so. (**2**) **Zone 2** **Sectorial information systems** – Systems of different public organisations providing and consuming data. They communicate through the State Administration Integrated Electronic Communications Network. (**3**) **Zone 3** **National Data Centre** (NDC) – This area includes the central e-government building blocks.

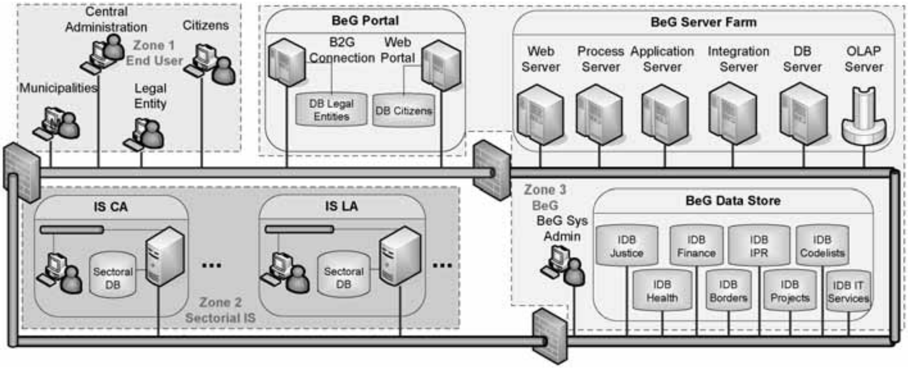


Fig. 3 Bulgarian e-government model 2010

NDC supports all major functions required to operate BeG. NDC is composed of two clouds – one for the central and one for the local administration. NDC is comprised of the following main components - security management, portal, communication management, process management, service management, management of structured and unstructured data. NDC main functions are supported by a Disaster Recovery Center.

6 Conclusion

Based on BeG development context presented therein before, impediments to the rapid realisation of e-government in Bulgaria are identified. The presented comparative analysis shows both differences in the development approach of the e-government systems and tangible common problems faced by all organisations developing these systems. Based on the results of the comparative analysis are defined important BeG requirements, governance scheme and development tools. Proceeding from these requirements BeG development concept is defined.

References

Australian Department of Finance and Deregulation (ADFD) (2013). Whole-of-Government Common Operating Environment Policy. http://www.finance.gov.au/files/2013/01/WofG-COE-Policy.pdf

Australian Government Information Management Office (AGIMO) (2011). Australian Government Architecture Reference Models. http://www.finance.gov.au/sites/default/files/AGA-RM-Final-v3.0-July-2013.pdf

Austrian Federal Chancellery (2014). Administration on the Net. The ABC guide of eGovernment in Austria. Kny & Partner, Druck & Verlagsproduktionen, Vienna.

Council of Ministers (CoM) (2016). Public administration status report 2015 http://www.government.bg/fce/001/0211/files/2806_Doklad_za_administraciata_2015_final-28062016.pdf p.41 (in Bulgarian)

Dutch Government (2010). Dutch Government Reference Architecture Strategy Supplement. https://www.digitaleoverheid.nl/images/stories/architectuur/nora_maart%202010-eng.pdf

European Commission's National Interoperability Framework Observatory (NIFO) (2016). eGovernment Factsheets. https://joinup.ec.europa.eu/community/nifo/og_page/egovernment-factsheets

Finnish Population Register Centre (FPRC) (2016). National architecture for digital services in Finland. https://esuomi.fi/mdocs-posts/national-architecture-for-digital-services-3/

IBM. Rational Method Composer. http://www-03.ibm.com/software/products/en/rmc

New Zealand Government (2015). Government Enterprise Architecture for New Zealand Reference Models and Taxonomies. https://www.ict.govt.nz/guidance-and-resources/architecture/government-enterprise-architecture-for-new-zealand-framwork/gea-nz-reference-models-and-taxonimies/

Tinholt D. et al. (2015). eGovernment Benchmark report. European Commission Directorate General for Communications Networks, Content and Technology. https://ec.europa.eu/digital-single-market/news/eu-egovernment-report-2015-shows-online-public-services-europe-are-smart-could-be-smarter

United Nations (UN) (2016). E-Government Survey. E-Government in support of sustainable development. http://workspace.unpan.org/sites/Internet/Documents/UNPAN96407.pdf

Vassil K. (2015). World development report 2016 Estonian e-Government Ecosystem: Foundation, Applications, Outcomes. http://pubdocs.worldbank.org/en/165711456838073531/WDR16-BP-Estonian-eGov-ecosystem-Vassil.pdf

Bulgarian e-Government Information System Based on the Common Platform for Automated Programming – Technical Solution

Ivan Stanev, Maria Koleva

Faculty of Mathematics and Informatics, University of Sofia St. Kliment Ohridski
5 James Bourchier blvd., 1164 Sofia, Bulgaria
instanev@fmi.uni-sofia.bg, mkoleva@fmi.uni-sofia.bg

Abstract. Bulgarian e-government information system (BeG) development stages are detailed. BeG-1 technical solution is described and analyzed. Improvements for BeG-2 are suggested. Architecture for BEG-2 is developed based on cloud computing, Service Oriented Architecture (SOA) and knowledge based automated software engineering (KBASE) complemented with customized components, including kernel, internal e-government components and external interfaces. BeG-2 results are presented and expected benefits analyzed. Enhancements for BeG-3 are proposed.

Keywords: SOA, Cloud computing, Knowledge based automated software engineering, Common platform for automated programming, e-government.

1. Introduction

BeG important problems are identified in [Stanev, 2016]. A comparative analysis of some of the leading technical solutions in the area of e-government are presented. On the basis of this analysis are defined important objective and subjective BeG problems. Defined are the general requirements and the concept for BeG development.

BeG development concept is detailed further on through its three stages performed in the following time periods: (1) Stage 1 - 2002-2010; (2) Stage 2 - 2010-2013; (3) Stage 3 - 2014 - 2020. For each stage is described the software architecture as well as the strengths and weaknesses of the proposed solution.

2. BeG Stage 1

2.1. BeG Stage 1 Software Architecture

The BeG e-government model 2007 is presented in Fig. 1. This model is laid down in the e-Government Act (EGA) adopted in mid-2007. The EGA arranges the activities of the organizations providing public services in relation to working

V. Dimitrov, V. Georgiev (Editors): ISGT'2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

with and exchanging electronic documents as well as providing electronic administrative services. EGA introduces **three important principles**: (1) personal data must be collected once and used many times; (2) citizens must be notified ex-officio; (3) documents must be automatically submitted.

Procedures were created for: (1) registration of data objects and e-services; (2) certification of administrative information systems (AIS); (3) realization of the Electronic Documents Exchange Environment (EDEE).

The technical prerequisites necessary for the functioning of this model were put in place by the Ministry of Transport, Information and Communication Technologies in the terms prescribed by the EGA.

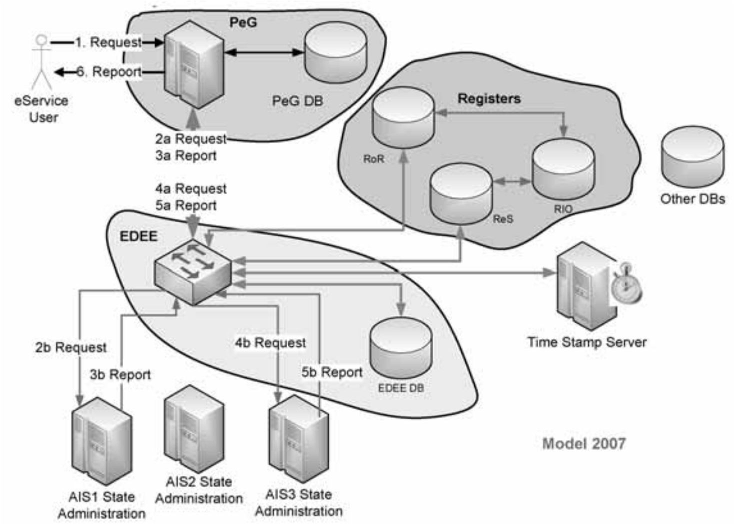


Fig. 1 Bulgarian e-government model 2007 [MTITC, 2010]

For the purpose of ensuring interoperability of the electronic document exchange between administrative bodies the following registers were developed: (1) The **Register of standards** (RS) contains technical standards to be applied by administrative bodies for the provision of e-services, interoperability, information security and automated exchange of information and documents between them. (2) The **Electronic Services Register** (ReS) contains description in XML format of all electronic administrative services, provided through the Electronic Documents Exchange Environment (EDEE). (3) The **Register of information objects** (RIO) contains standardized descriptions in XML format of information objects (terms, nomenclatures, documents, etc.), which are collected or created by any administrative body. (4) The **Register of registers** (RoR) contains data for all registers, lists, etc. maintained by the public authorities, provides a unique index for each data set, and supports unified definitions of the stages of the administrative services.

The **e-Government portal** (PeG) comprises a catalogue of public services provided by the central state administration and enables citizens and businesses to obtain online information about public services, as well as forms to download. Most of the available services are information services. There is a limited number of services requiring electronic signature.

The **electronic signature** supported by a qualified certificate issued by a registered certification service-provider is used as a general instrument for authentication of citizens and representatives of the legal entities in their communication with the state authorities. The **Uniform Citizen Number** is extracted from the user's certificate for electronic signature for identification purposes.

Electronic Documents Exchange Environment was developed in order to allow standardized and secure exchange of structured and unstructured documents among the information systems of all e-government participants.

In accordance with the EGA all government organizations ought to certify their information systems to use the EDEE.

Although the 3 main components of this model, namely PeG, EDEE and the e-government registers, were up and running as prescribed by the EGA, no one document is exchanged through the EDEE in production environment.

2.2. BeG Stage 1 Strengths & Weaknesses

Important **BeG strengths** for Stage 1 are the following: (**1**) Single Communication Point introduced; (**2**) Common BeG service bus realized; (**3**) Common BeG central registers created.

Key **BeG weaknesses** are the following: (**1**) No active services at the BeG Portal; (**2**) No documents exchanged through the common BeG service bus; (**3**) Not working interoperability concept; (**4**) No BeG concept and Common Software Architecture.

3. BeG Stage 2

Model 2007 was built following the EGA principles but they were not implemented into practice.

Following the BeG development concept proposed in [Stanev, 2016], Stage 2 technical solution shall enable BeG to:

- Revoke the existing semantic interoperability concept, which is based on registration of e-services and data objects in the interoperability registers, as well as the term "administrative information system";
- Develop a new concept for semantic interoperability based entirely on standardized exchange of information between eGRs (e-government registers, [Stanev, 2016]);
- Remove the need to register e-services and data objects, as well as to certify

AIS. Dissolve the relevant Registration authority;

- Develop a concept providing for the publication of new e-services and data by the individual administrations. The technical parameters and access rights to proprietary services and data shall be defined. These e-services and data shall be available to partner organizations through official channels automatically and in real time;
- Develop a common agreement between owners and users of e-services and data, providing for their rights and obligations and ensuring the absence of further formalities and actions with the exception of: (1) testing the integration with new systems; (2) one time registration of new users before the first use of the relevant resource; (3) defining roles and rights of new users;
- Implement a new concept for the exchange of information between administrations based on the new enterprise service bus to replace the existing Electronic Documents Exchange Environment and allowing synchronous data exchange to be the norm.
- Introduce the Rational Unified Process (RUP) framework for all units of the state administration.

3.1. **BeG Stage 2 Software Architecture**

BeG model 2010 is presented in Fig. 2.

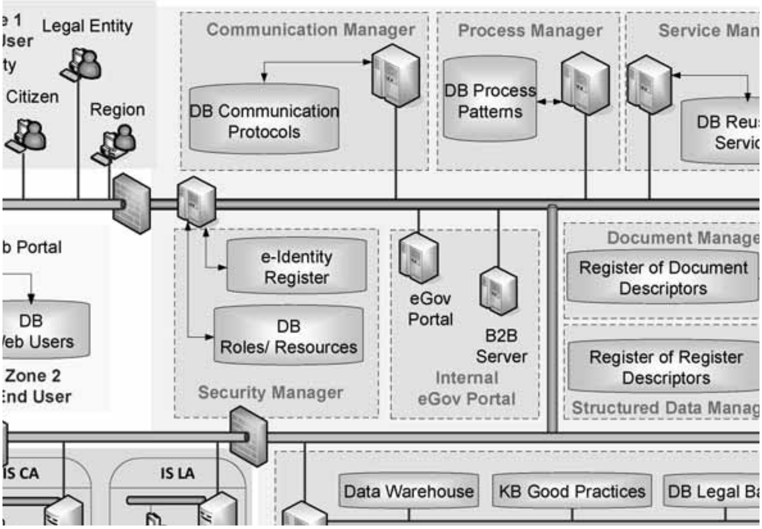


Fig. 2 BeG model 2010 – National Data Centre Components[1]

[1] Abbreviations: CA – Central Administration, DB – Data base, IS – Information system, KB – Knowledge Base, LA – Local Administration, NDC - National Data Centre

BeG model 2010 components are described below.

The **Web Portal** provides complete and up-to-date information to end users as well as a platform and a single access point for digital services. End-users are identified, apply for e-services and receive results. In addition, support functions are provided to assist users in applying for e-services such as: Access to support materials and content; Possibility for interactive Q&A sessions; User profile management; Search, etc. Except for a standard Web browser, the portal shall be available for mobile devices.

The **sectoral ISs** are systems of different public organizations providing and consuming data. They implement the business logic for the provision of the different e-services. They represent the end point where e-service requests arrive and the starting point where the employees of the relevant administration begin processing the request. Through different application and integration interfaces, the IS should exchange data with the Communication manager. They communicate through the State Administration Integrated Electronic Communications Network.

All external systems will communicate with NDC components through the **Communication Manager**. For this purpose, the manager will provide configurable XML Web services. The Communication Manager is also responsible to validate the incoming data. It is used mainly for:

- Implementing the principle of "official notification" – after changing the data in the IS of the primary data administrator all interested parties are "notified" by calling web services;
- Implementing the principle of "single collection and creation of data" - when additional data for the subject of the e-service is needed, the e-service provider addresses the respective primary administrator to retrieve the data;
- Provision of internal e-services - by calling web services, in which the data is packaged (I / O) for the respective internal e-service.

The **Process Manager** manages complex business processes in the exchange of messages between different components (internal and external) of the central integration platform. It manages the transformation of a non-formal specification into a complete and consistent formal specification of business processes to be realized. It manages the interpretation of the specified process by the Process Server. The Process Manager works with services and interfaces available from the Communication Manager. Used mainly for: Implementation of the e-services provided through the web portal or located in the NDC infrastructure; Orchestration of internal NDC processes.

The **Service Manager** is used to manage the processes of services description, update, deletion and interpretation. It includes standard descriptors (e.g. service, endpoint, binding, interface, operation, data types in the Web Services Description Language specification). Upon change in a service, the Service Manager notifies the Process Server. A message is subsequently sent to all users of the service. The Service Manager contains a repository of metadata for the processes published by the process server and the web services virtualized by the service bus. The information that should be managed by the manager: Integration interface of the process/service; Rights and access policies to the process / service; Dependence of the process/service on other components; History of changes of the process/service.

Technical metadata repository should be integrated with the interoperability register containing e-services business metadata.

The **Security Manager** consists of components that enable centralized identification and authorization of e-service users. Key element of the Security Manager is the central repository of all registered users. The information in this repository is enriched and checked against additional external sources. The identification component supports different identification schemes (user name and password, eID, etc.). Users are authenticated by the Security Manager for the purposes of e-services delivery. The Security Manager provides authorization functions to modules, processes, services and data based on a roles – resources – permissions matrix that presents the rights of each role with respect to every object in the system. Working with certificates (validation and optional issuing) requires a Public Key Infrastructure in order to manage the corresponding cryptographic artefacts in a secure environment.

The **Internal eGov Portal** is the only BeG entry point which implements the single window concept. The portal is responsible to organize user access to modules inside the NDC platform, to enable users to describe what work they require to be performed by the system and to provide them with the results. The interaction between BeG users and NDC is performed trough the eGov Portal or between BeG user information system and NDC in B2B (Business To Business) communication.

The **Document Manager** manages the processes of creating, editing, versioning, and deleting cases from the central case catalogue, and the central case library (CLib, see DB Documents below). It is responsible to make available to users the cases included in the catalogue of CLib the Manager is working with. It manages requests from both internal and external users to access the central and partner CLib. Implements document management workflows involving more than one administration.

The **Structured Data Manager** maintains an integrated catalogue of registers and executes queries to the catalogue. Provides tools for automatic exchange of information between registers.

The **Database Manager** is responsible for management and monitoring of the common databases of BeG. The Data Warehouse is a central repository of data from the NDC and partner information systems. It is used for creating ad-hoc and canned analytical reports for knowledge workers, organization managers, etc. In KB Good Practices are gathered good practices algorithms such as a customized software development process. DB Legal Base provides national legal and legislative information. This may include information on the lifecycle of a legislative proposal, preparatory acts, the Official Journal, consolidated acts, case-law, international agreements, standards etc. DB Documents represents the central case library where are stored documents processed by the Template Information Systems (TIS) for document management. TIS are designed for exchange of specific data from homogeneous sources. The TIS concept shall replace the development of a large number of similar information systems requiring significant resources for realization. E-government registers (eGRs) data is stored in DB Registers. eGRs represent an excerpt from the registers of the primary data administrator for which eGRs users have legally justified access right. In DB Code lists are stored code lists (CL) developed and maintained by their relevant primary administrator and used by multiple organizations. This DB may include code lists of countries, cities, economic activities, education and training areas, etc.

3.2. **BeG Stage 2 Results**

The next figure demonstrates the technical solution developed as a result of two projects performed in the period 2010 – 2013 and co-financed by the Administrative Capacity Operating Program. Components numbered 31x are developed under Project K10-31-1 / 07.09.2010 "Extension of administrative service delivery through electronic means" (K10311). Components numbered 32y are developed under Project K11-32-1/ 20.09.2011 "Better administrative service delivery through further development of the central e-government systems" (K11321).

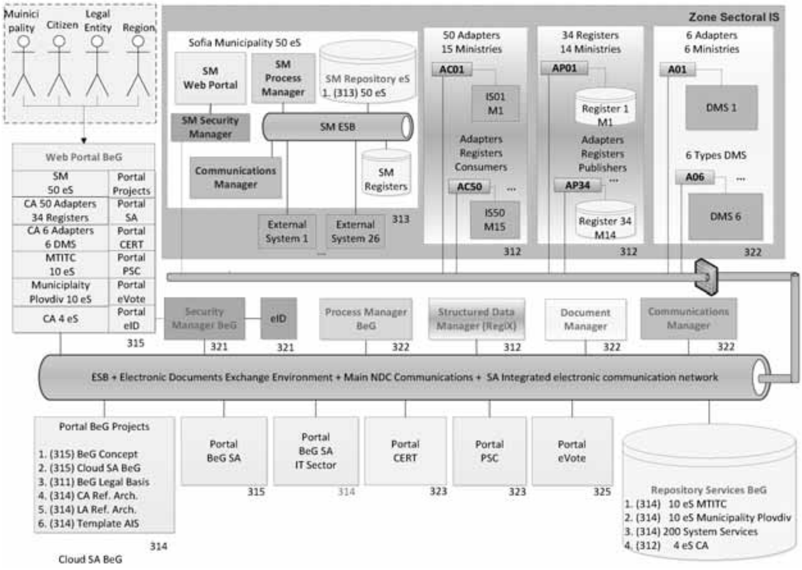


Fig. 3. Bulgarian e-government model 2010 – Results

Stage 2 projects build on the existing e-government infrastructure, namely State Administration (SA) electronic communication network, existing National Data Centre, Electronic Documents Exchange Environment. The NDC was challenged to meet the needs of Stage 2 developments. Under lot 326 were built three virtualization platforms. Staging and production environments were prepared for the systems developed under projects K10311 and K11321, all existing systems were migrated.

The e-Government Act, related implementing provisions and instructions are revised with focus on e-services under lot 311.

As a result of lot 314 the reference architecture models of e-government, central administration and local administration are developed. .Net and JEE prototypes of the document exchange between administrations is developed.

Primary data administrators at central and local level and services they provide to citizens, businesses or other administrative bodies are analyzed. Requirements for development of Government-To-Government (G2G), Government-To-Citizens (G2C) and Government-To-Business (G2B) e-services are gathered. The priority requirements are realized under three lots.

312 – 34 registers of the central administration (CA), i.a. Commercial register, Property register, Cadaster register, Citizens personal data register, Criminal record register, etc., are open to provide legally authorized data to partner organizations, mainly for the purposes of complex e-services. 50 adapters to consumer registers are developed. The Structured Data Manager manages a catalogue of registers and handles the interaction between adapters to registers. Adapters to published registers (APx) are installed in the data center of the corresponding register owner (Mx, ministry). On request by the Structured Data Manager, the adapter communicates with the information system (IS) of the register owner, filters the extracted data in accordance with the specified authorization rules and transfers the data to the requesting organization via Enterprise Service Bus (ESB). Adapters to consumer registers (ACx) are installed in the data center of the requesting organizations.

313 – A SOA platform for Sofia Municipality (SM) is developed based on open source tools and international standards. 50 reusable e-services are developed.

314 – 20 services for the central and local administrations are developed.

The Document Management System (DMS) developed under lot 322 manages central catalogue and repository of documents and cases as well as the provision of requested documents and cases from its own and from partner repositories. Provides traceability of electronic services implementation in each participating administration. This system could be used by any administration that does not have its own document management system.

Under lot 321 is developed the Bulgarian electronic identification (eID) system in accordance with EU regulations. The system covers all processes related to eID issuing and validation. 5000 pilot eID cards are issued to civil servants, journalists, representatives of universities and the private ICT sector.

G2C and G2B e-services are accessed via the Web portal. Under lot 315 the existing portal is upgraded with functionality for portal customization, reusability of information, and integration, thus enabling easy administration and maintenance via user interface. The e-government portal serves as a one-stop shop comprising a repository of public services provided by the State Administration (SA).

The Web portal provides access to a number of internal portals. The Interoperability of information systems portal (Portal BeG SA IT sector) developed under lot 315 maintains a database of documents and materials related to interoperability and is designed for developers of information systems and IT specialists.

CERT (Computer emergency response team) system developed under lot 323 provide functionality for registration and follow-up of computer security incidents, installation base maintenance, and security vulnerabilities management.

The Point of Single Contact (PSC) developed under lot 323 is an extension of the national e-government portal allowing citizens and business to find out about the formalities that apply to services and to complete the administrative procedures online.

The eVote portal developed under lot 325 is used for online participation of citizens and business in government decision making process.

G2G e-services implementation relies on the Enterprise Service Bus. The ESB realized under lot 322 enables a unified and flexible approach for the integration between heterogeneous systems and provides an integration environment that is not dependent on the platforms and architectures of the various systems that will integrate now or in the future. Within the project was created an UDDI (Universal Description, Discovery, and Integration) register of publicly available web services.

Bulgarian e-government (BeG) development Concept is prepared under lot 315. It is based on SOA, cloud technologies, open instruments and international standards.

3.3. **BeG Stage 2 Strengths & Weaknesses**

More important **BeG strengths** of this period are: (**1**) Introduced Single Window technology; (**2**) Realized Enterprise Service Bus; (**3**) Realized Concept for Centralized Identification and Authorization management; (**4**) Realized Interoperability Concept based on State Registers; (**5**) Introduced Service Oriented Architecture; (**6**) Established SA Centralized service repository; (**7**) Realized Concept for Template IS (e.g. Centralized Document Management System, etc.); (**8**) Created Cloud and SOA of Local State Administration.

Key **BeG weaknesses** are: (**1**) Insufficient BeG infrastructure; (**2**) Low level of communication infrastructure; (**3**) Low level of SA IT expertise; (**4**) Insufficient SA business processes restructuring.

4. **BeG Stage 3**

4.1. **BeG Stage 3 Software Architecture**

Based on the analysis of Stage 2 results, the requirements and the concept defined in [Stanev, 2016], as well as the results presented in [Stanev, 2015-2], BeG Stage 3 Software architecture is suggested below in Fig. 4.

The main technical objective of Phase 3 is the development of a Common Platform for Automated Programming (CPAP), based on the following principles:

- CPAP components are divided in the following layers: (1) physical layer; (2) virtualization layer; (3) system administration layer; (4) components generator layer; (5) business components layer of; (6) layer for integration with external systems. Layers communicate through a standard and a semantic Enterprise Service Bus.
- CPAP components are designed to build: (1) embedded systems - for collecting information; (2) information systems – for information processing; (3) knowledge based systems - for automated processing of large data sets, for generating new components and for monitoring and reconfiguration of CPAP.
- CPAP models in P11 are built by the (1) Embedded Systems Manager, (2) Information Systems Manager, and (3) Knowledge Base Manager. Based on the models, they generate in collaboration three meta-models in P11, namely the IS Model, KBASE Model and Embedded Systems Model, as well as the integration definitions in P20.
- The static and dynamic behavior of the components are formally described by P20 definitions using one or more languages and standards such as BPMN (Business Process Model and Notation, [OMG, 2014]), UML (Unified Modeling Language, [OMG, 2015-2]), SysML (Systems Modeling Language, [OMG, 2015-1]), Net [Stanev, 2001].
- The static and dynamic behavior of the systems are formally described by P20 definitions using one or more languages and standards such as BPMN, UML, CMMN (Case Management Model and Notation, [OMG, 2016-1]), DMN (Decision Model and Notation, [OMG, 2016-2]), Net, OWL (Web Ontology Language, [W3C, 2009]).
- Components used at CPAP central level are generated by: (1) Embedded Systems Manager, (2) Information Systems Manager, and (3) Knowledge Base Manager, using P20 integration definitions, P10 repository of services and components, and P11 knowledge.

- The work between CPAP and its external users is organized by Template Information Systems (TIS) generated by the TIS Manager. The TIS Manager is managing the IS Manager, KBASE Manager, and Embedded Systems Manager for the purposes of collecting data from external systems. The TIS Manager manages the generation of the TIS meta-model in P11 and TIS integration definitions in P20 for internal and external use. The TIS Manager subsequently generates the TIS in collaboration with the Embedded Systems Manager, Information Systems Manager, and Knowledge Base Manager.
- Each TIS may be built as a combination of one or more CPAP components and could perform one or more of the following tasks: (1) provide standardized and controlled exchange of information with partner information systems operating in their own environment, external to CPAP; (2) provide client-server mode functionality between the CPAP central level (TIS Server) and external users (TIS client); (3) to manage data exchange at different levels as follows: Central CPAP level (TIS Server Level 1), central level of an external user (having simultaneously the roles of TIS Client Level 1 and TIS Server Level 2), and local level of an external user (TIS Client Level 2); (4) manage the operation of a centralized TIS open for direct access to external users.

CPAP architecture is composed of six layers. **L1 Infrastructure Layer** organizes and manages the hardware and communication infrastructure processes at physical level (P01 Hardware), operating system level (P02 Real OS), and using virtualization tools (P03 Virtualization). **L2 Cloud Layer** automatically organizes the execution of requested work in the cloud environment. This is performed under the control of the virtual operating system (P04 Virtual OS), using scaling mechanisms (P05 Cloud instances), and by allocating the required physical and virtual resources (P06 Cloud cartridges). **L3 SOA Layer** manages user assignments execution at service level (through P07 Application Servers) and process level (through P08 SOA Servers). **L4 Ontology Layer** contains data (P09 Data), components and services (P10 Repository), and knowledge (P11 Knowledge) needed to generate required software products using BPMN and/ or UML specifications, graphical interfaces, natural language, etc. Models included in P11 are subject to various automated verifications such as verification and modification based on predefined standardized knowledge bases [Stanev, 2012-2], business process model quality assessment [Krastev, 2015], business process semantic correctness evaluation [Shahinyan, 2014]. **L5 Business Layer** consists of a rich set of ready-made tools that are provided on demand by the user to assemble his software product. Components in L5 are organized in three groups of packages. The first group consists of components for <u>development management</u> (P12 Development management) responsible for managing the process of creating new software products in the context of CPAP and updating them on demand (possibly in real time) ([Stanev, 2012-2], [Stanev, 2012-1], [Krastev, 2017]). The second group consists of components for runtime management (P13 Runtime management, P19 Presentation management). The third group consists of components for business logic management (P14 Organization Management, P15 Document Management, P16 Activity Management, P17 Collaboration Management, and P18 User Support).

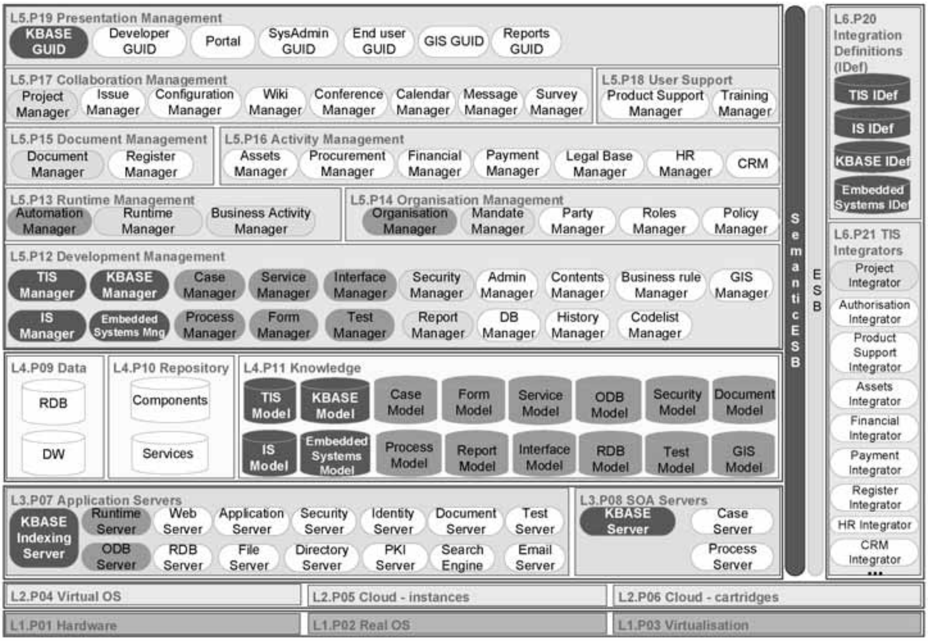


Fig. 4 BeG Model 2020 – Cloud and KBASE Solution

L6 Integration Layer includes integration definitions and products for integrating systems, processes and services developed within the organization, from partners or third parties. Package P20 is composed of integration definitions (IDef) for TIS, Information Systems (IS), KBASE, and Embedded Systems. IDef represent integration conventions (e.g. communication protocols, XML schema definitions, dialogue sequence, information semantics, etc.) generated by P12 components. Integration could be performed at architectural or execution level. KBASE integration is performed in order to integrate knowledge from various external systems. Embedded systems integration is performed at architectural [Patias, 2015-1] or at business information flow level [Patias, 2015-2]. Integration of components and services for information processing is performed based on IS IDef. TIS IDef for a specific domain area are generated using IDef for IS, KBASE and Embedded systems. P21 TIS Integrators comprise TIS server components (TIS Server Level 1).

CPAP is composed of four types of components located in different CPAP layers. These four types are as follows: (1) Components for classic programming (CCLP); (2) Components for combined programming (CCP); (3) Components for automated programming (CAP); (4) Components for ontology programming (COP).

Components for **classic programming** (in *white fill color* in Fig. 4) allow the realization of classic programming techniques related to multi-tier architectures, SOA ([OASIS, 2006]) and CC ([Mell, 2001], [Liu, 2011]) and are of little interest as regards programming automation. However these components are the minimal body of technical means sine qua none standard software products would not operate. Components for **combined programming** (in *turquoise fill color* in Fig. 4) represent techniques for classic programming enriched with automation elements, e.g. dynamic reconfiguration of data structures or computing process management based on business rules or business process specifications, etc. Components for **automated programming** (in *light blue fill color* in Fig. 4) represent automation tools based on the interpretation of formal models (e.g. direct generation of software products from UML, BPMN, EPC, fourth generation languages, formal methods, etc.). Components for **ontology programming** (in *dark blue fill color* in Fig. 4) provide for automation through knowledge interpretation, learning and self-learning (e.g. generate software based on ontology descriptions, fuzzy interpreters of incomplete and inaccurate specifications, code generators working with natural language specifications, etc.).

4.2. **BeG Stage 3 Roadmap**

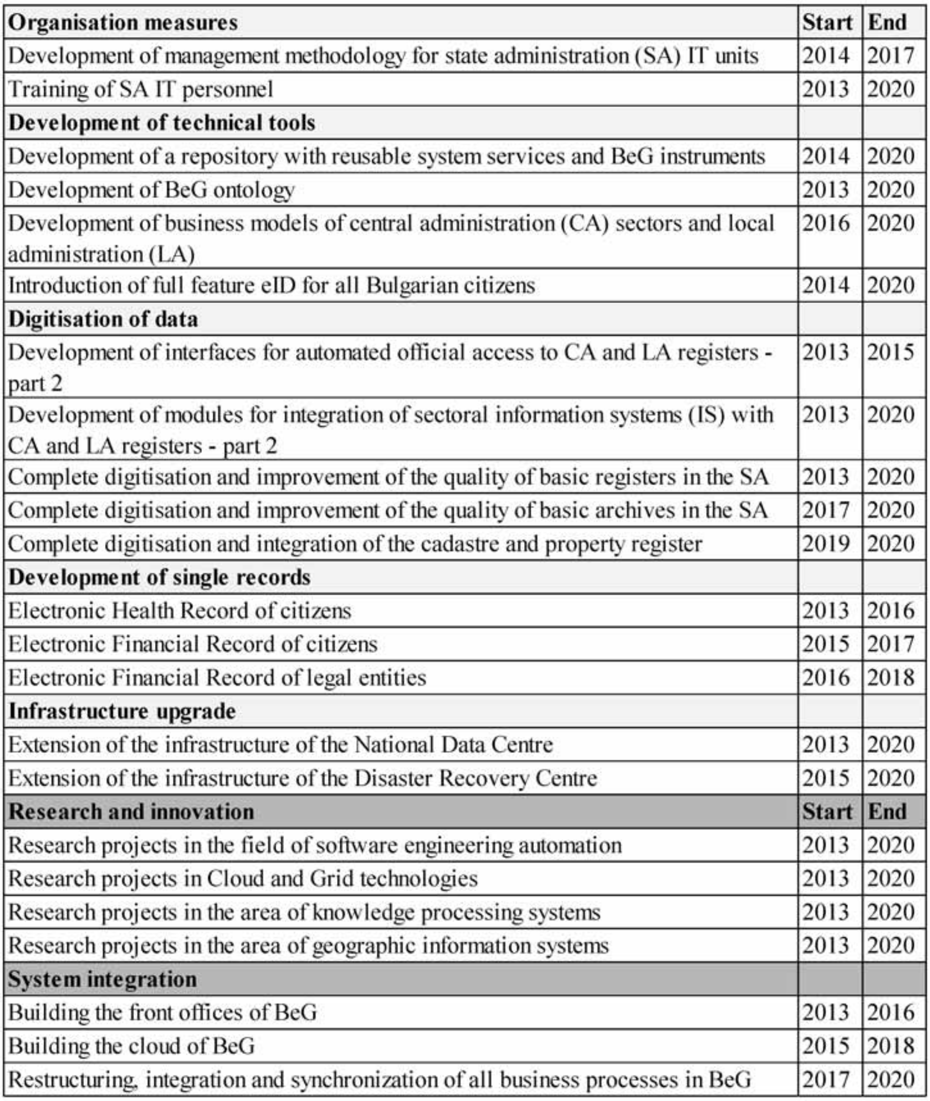
The Roadmap for realization of BeG Stage 3 is composed of interrelated projects focused on BeG organizational and technological aspects. These projects with their estimated implementation period are presented in Fig. 5 below.

Important measures for Stage 3:

- Organizational measures include the development of **IT processes management** methodology for SA units covering IT systems planning, development and maintenance lifecycle as well as provision of the necessary training.
- Development of a **repository with reusable services** to be freely available to administrations and their future IT partners. The system services may include infrastructure management services, utility services, financial services, universal report generators, standardized dialogue services, statistical data processing services, etc. Reusable business services shall be identified during business model development by SA sectors and LA. They may be services available in the repository. Further steps towards real restructuring

of the sectoral business processes shall be carried together with legal base harmonization

- **BeG ontology model** shall be developed with priority on standardized knowledge bases (including classification and organizational schemes, performed processes, data processing techniques, statistical processing techniques, etc.), databases and procedures for their use standardizing and describing the work in the administration.



Organisation measures	**Start**	**End**
Development of management methodology for state administration (SA) IT units	2014	2017
Training of SA IT personnel	2013	2020
Development of technical tools		
Development of a repository with reusable system services and BeG instruments	2014	2020
Development of BeG ontology	2013	2020
Development of business models of central administration (CA) sectors and local administration (LA)	2016	2020
Introduction of full feature eID for all Bulgarian citizens	2014	2020
Digitisation of data		
Development of interfaces for automated official access to CA and LA registers - part 2	2013	2015
Development of modules for integration of sectoral information systems (IS) with CA and LA registers - part 2	2013	2020
Complete digitisation and improvement of the quality of basic registers in the SA	2013	2020
Complete digitisation and improvement of the quality of basic archives in the SA	2017	2020
Complete digitisation and integration of the cadastre and property register	2019	2020
Development of single records		
Electronic Health Record of citizens	2013	2016
Electronic Financial Record of citizens	2015	2017
Electronic Financial Record of legal entities	2016	2018
Infrastructure upgrade		
Extension of the infrastructure of the National Data Centre	2013	2020
Extension of the infrastructure of the Disaster Recovery Centre	2015	2020
Research and innovation	**Start**	**End**
Research projects in the field of software engineering automation	2013	2020
Research projects in Cloud and Grid technologies	2013	2020
Research projects in the area of knowledge processing systems	2013	2020
Research projects in the area of geographic information systems	2013	2020
System integration		
Building the front offices of BeG	2013	2016
Building the cloud of BeG	2015	2018
Restructuring, integration and synchronization of all business processes in BeG	2017	2020

Fig. 5 BeG Stage 3 Roadmap

- Each Bulgarian citizen shall be provided with **electronic identification (eID) card** containing electronic identity, electronic signature and e-wallet. The eID could be used as a national ID card for travel within the EU, as a national health insurance card, as proof of identification when logging into bank accounts, for e-vote, for accessing government systems to check one's medical records, file taxes, etc. The eID card shall be issued by the eID governance authority or by authorized eID administrators [EIDA, 2016] on various means, i.a. bank card, SIM card, etc. For the purposes of eID issuing and validation a centralized register is maintained and used by all relevant authorities. eID authentication for the purposes of e-services delivery is provided by the eID validation authority or by authorized organizations.
- Extend the development of interfaces for **access to the registers of CA and LA** as well as modules for integration of sectoral IS to the published registers to allow the administrations to automatically exchange data they have already collected. This effort shall be supported by the full **digitization** and improvement of **the data quality** of the main registers and archives in the administration. Tools that support data cleaning, data enrichment, data integrity and data quality assurance shall be put in place.
- **Infrastructure development** shall be aligned with the development and deployment of new e-services for the end users (companies and citizens) and for the partner organizations (e.g. central and local administration) resulting in increased infrastructure load. BeG systems reliability, security, performance and supportability shall be increased to ensure fault tolerance, load balancing, security control, real-time monitoring. The development of **BeG Cloud** solution shall be initiated and build on what has been achieved. This will streamline the work on integration of the sectoral information systems into a unified e-government system.
- **Research and development** projects shall be undertaken in the areas of **automated software engineering** (development of automated software generation tools, automated testing tools, infrastructure quality of service tools, tools for automated developing of self-organizing systems, etc.), **knowledge management** (development of new ontologies for CA and LA, intelligent search engines, tools finding solutions based on incomplete and inaccurate information, multilingual systems management, etc.), **GIS management** (intelligent image recognition, automated linking of graphical and text information, automated synchronization and harmonization of combined graphical and textual information, etc.).
- **BeG** shall be **brought closer** to people that are disconnected by choice or necessity by establishment of e-government front-offices in the existing postal offices located in underdeveloped regions.

5. **Expected Results**

For the successful implementation of BeG Stage 3 it is necessary to: (1) elaborate a new e-Government Act (EGA) and make the necessary amendments to the Law on Electronic Document and Electronic Signature (partially implemented with the amendments in the e-Government Act from 20.05 and 01.07.2016), as well as align the sectoral legal base with EGA; (2) build a national data center (in progress); (3) build a cloud computing platform (in progress); (4) develop a common procedures for access to electronic data and services (in progress); (5) establish a national digitization center for the CA and LA, as well as launch projects to improve the quality of data in the registers of CA and LA.

If the above prerequisites are met the following is planned to be achieved during Stage 3:

(1) Deploy 150-180 new electronic services per year (in case CA sectors legal base is aligned, business processes are restructured and core registers are fully digitized and error-free, see prerequisites №№1, 4, 5);

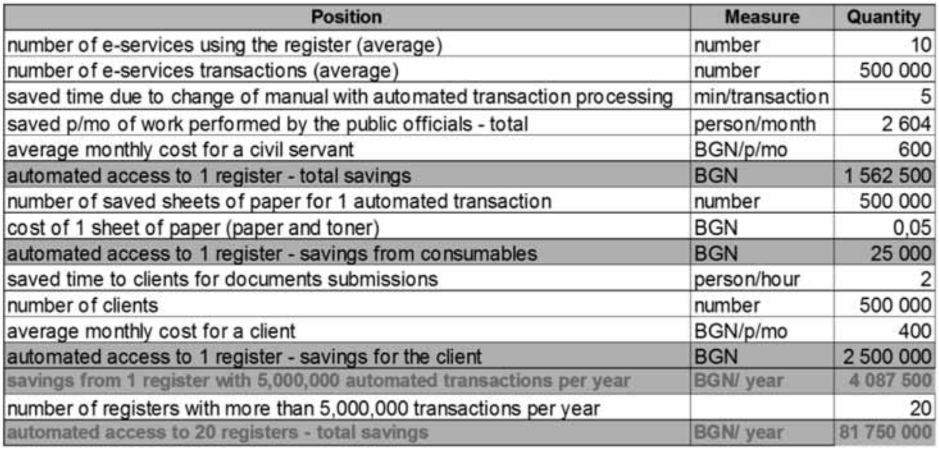
(2) Publish immediately in all possible formats the data from the Open Data initiative (in case the registers of the respective administration are included in the Register of Registers of the Structured Data Manager component and there is a legal base providing for the publication of the relevant data, see prerequisites №№1, 2, 3);

(3) Deploy 3 to 5 electronic services per month for all local administrations (if LA cloud is upgraded, see prerequisite №3, 4);

(4) Digitization of 10-15 registers per year (after the establishment of a National digitization center, see prerequisite №5).

The following table shows that it is possible to directly save about BGN 80 million from the state budget annually for each 10 new electronic services with more than 500 000 customers per year if put into operation following the concept presented hereinbefore.

Table 1. Expected financial benefits.



Position	Measure	Quantity
number of e-services using the register (average)	number	10
number of e-services transactions (average)	number	500 000
saved time due to change of manual with automated transaction processing	min/transaction	5
saved p/mo of work performed by the public officials - total	person/month	2 604
average monthly cost for a civil servant	BGN/p/mo	600
automated access to 1 register - total savings	BGN	1 562 500
number of saved sheets of paper for 1 automated transaction	number	500 000
cost of 1 sheet of paper (paper and toner)	BGN	0,05
automated access to 1 register - savings from consumables	BGN	25 000
saved time to clients for documents submissions	person/hour	2
number of clients	number	500 000
average monthly cost for a client	BGN/p/mo	400
automated access to 1 register - savings for the client	BGN	2 500 000
savings from 1 register with 5,000,000 automated transactions per year	BGN/ year	4 087 500
number of registers with more than 5,000,000 transactions per year		20
automated access to 20 registers - total savings	BGN/ year	81 750 000

6. Conclusions

During BeG Stage 2 implementation the following solutions are introduced: (1) a concept for BeG development is proposed, based on cloud technologies; (2) a software architecture based on services is developed and implemented as a common service bus; (3) a technological platform for the realization of BeG software architecture is developed, composed of tools for virtualization and automated management of the scalability, reliability and security of BeG; (4) centralized identification and authorization with distributed management are introduced; (5) the “one-stop-shop” principle is introduced; (6) the interoperability mechanism based on the state administration registers is introduced; (7) ontologies were introduced to solve the problems of semantic interoperability.

Based on Stage 1 analysis, Stage 2 implementation and achieved results, Stage 3 design and expected results, and the prepared BeG roadmap until 2020 the following conclusions could be drawn:

Problems solved: (1) Improvement of data quality; (2) Improvement of interoperability; (3) Improvement of standardization.

Achieved BeG parameters improvement: (1) Decrease of time for software development; (2) Decrease of time for incorporation in BeG Software of Legal base changes; (3) Decrease of new software cost; (4) Improve software flexibility and reusability.

Improved quality of service: (1) Considerable decrease of cost for state administration; (2) Improvement of quality of services for citizens.

References

Electronic Identification Act (EIDA) (2016). http://www.lex.bg/bg/laws/ldoc/2136822116

Krastev E. (2017). Mathematical Model for Motion Control of Redundant Robot Arms. In Proceedings of the 19th International Conference on Mathematical and Computational Methods in Science and Engineering (Berlin, Germany, 2017).

Krastev E., Shahinyan K. (2015). Computer Assisted Quality Assessment of a Set of Business Process Models. In Proceedings of the 9th IEEE European Modelling Symposium of Mathematical Modelling and Computer Simulation (Madrid, Spain, 2015).

Liu F., et. all. (2011) NIST Cloud Computing Reference Architecture. Gaithersburg: National Institute of Standards and Technology Special Publication 500-292. US Department of Commerce. P. 35.

Mell P., Grance T. (2011). The NIST Definition of Cloud Computing. Gaithersburg: National Institute of Standards and Technology NIST Special Publication 800-145 US Department of Commerce. P. 7.

Ministry of Transport, Information Technology and Communications (MTITC) (2010). General Strategy for E-Government in the Republic of Bulgaria. http://www.strategy.bg/StrategicDocuments/View.aspx?lang=bg-BG&Id=662

Object Management Group (OMG) (2014). Business Process Model And Notation (BPMN) v.2.0.2, http://www.omg.org/spec/BPMN

OMG (2015-1). SysML v1.4. http://www.omg.org/spec/SysML/

OMG (2015-2). Unified Modeling Language (UML) v2.5, http://www.omg.org/spec/UML/

OMG (2016-1). Case Management Model And Notation (CMMN) v1.1, http://www.omg.org/spec/CMMN

OMG (2016-2). Decision Model And Notation (DMN) v1.1, http://www.omg.org/spec/DMN

Organization for the Advancement of Structured Information Standards (OASIS) (2006). Reference Model for Service Oriented Architecture. OAZIS. Pp. 31.

Patias I., Georgiev V. (2015-1). Embedded Architecture of Tolls Collecting System. In Proceedings of the 9th International Conference Information Systems & Grid Technologies (Sofia, Bulgaria, 2015).

Patias I., Georgiev V. (2015-2). Traffic Prioritization System Based on Embedded Components. In Proceedings of the 9th International Conference Information Systems & Grid Technologies (Sofia, Bulgaria, 2015).

Shahinyan K., Krastev E. (2014). Semantic Correctness of a Set of Business Processes. In Proceedings of the 10th Conference on Computer Science and Education in Computer Science (Albena, Bulgaria, 2014).

Stanev I. (2001). Formal Programming Language Net. Part II - Syntax Diagrams. In proceedings of the CompSysTech’2001 (Sofia, Bulgaria, 2001).

Stanev I. (2012-1). Method for Automated Programming of Robots. In Knowledge Based Automated Software Engineering. Cambridge Scholars Press. Cambridge. Pp.67 – 85. 2012.

Stanev I., Grigorova K. (2012-2). KBASE Unified Process. Knowledge Based Automated Software Engineering. Cambridge Scholars Publishing. Cambridge. Pp. 1 – 19.

Stanev I., Koleva M. (2015-2). Common Automated Programming Platform for Knowledge Based Software Engineering. In Proceedings of the ICSII 2015 17th International Conference on Semantic Interoperability and Integration (Rome, Italy, 2015).

Stanev I., Koleva M. (2016). Bulgarian eGovernment Information System based on the Common Platform for Automated Programming – Requirements. In Proceedings of the 10th International Conference Information Systems & Grid Technologies (Sofia, Bulgaria, 2016).

W3C (2009). Web Ontology Language (OWL) v.2. https://www.w3.org/standards/techs/owl#w3c_all

Hierarchical clustering in evolutionary analysis – data preparation and validation

Dilyana Hadzhimateva, Irena Avdjieva*, Valeriya Simeonova, Dimitar Vassilev

[1] Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”, 5 James Bourchier blvd., Sofia 1164, Bulgaria
* Correspondong author: Irena Avdjieva <ijstojchev@fmi.uni-sofia.bg>

Abstract. Phylogenecic trees can be viewed as a form of hierarchical cluster which visualize the relationsips between genes, proteins or entire species that share a common origin. This study is concentrated on plant gene trees and describes the processing of publicly available phylogenetic data with authors’ own Python scripts and open-source software. Its main goal is to develop methodologies for reduction, regrouping, manipulation and identification of homology relationships and topology patterns within the dataset, thus providing better knowledge about the evolution of major phylogenetic traits in plants.

Keywords: evolution, hierarchical clustering, phylogenetic tree, homology, Python, Ensembl

1 Phylogenetic trees as hierarchical clusters

In biology, the evolutionary relationships between entitites (genes, species, or other taxonomic units) are graphically represented as a phylogenetic tree. It can be describes as an undirected, bifurcating graph in which any two entities (nodes) are connected by exactly one path (branch). All entities in a tree share a common ancestor and descent from it into distinct lineages. The reconstruction and analysis of phylogenetic trees are an essential part of paleobiology, systematics, phylogenomics and any other fields researching speciation, origin or inheritance. Traditional phylogenetics used to rely on morphological data, but today phylogenetic trees are reconstructed from the sequence alignment of analyzed genes or proteins with either distance matrix or parsimony algorithms. There are many methods available for reconstructing the phylogeny from nucleotide or amino acid alignments. They can be grouped according to the kind of data they use, discrete character states or a distance matrix of pairwise dissimilarities, and according to the algorithmic approach of the method, either using clustering algorithm or optimality criterion. An agglomerative clustering algorithm usually

V. Dimitrov, V. Georgiev (Editors): ISGT’2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

results in only one tree estimate and is preferred even for large datasets. The two main approaches - neighbour-joining and UPGMA, form the basis of many integrated software applicatons, such as ‘MEGA’ [Tamura 2014].

2 Phylogenetic trees as hierarchical clusters

Each method for constructing phylogenetic trees results in a tree written in either Newick, Nexus, and PhyloXML file format. They carry information about the structure of the tree and sometimes contain additional parameters (the lengths of the branches, the sequences of the leaves, the results of the comparison, etc.). These formats are readable by many applications for visualization and analysis of phylogenetic trees. Newick (Newick notation or New Hampshire tree format [Archie et al. 1986]) is a format for representing trees in linear form defined lengths of the ribs. The original version of the form was created in 1984 by Christopher Mitchum needs of PHYLIP (suite of software applications to work with phylogenies). Newick format serves as a standard included in most software packages for constructing phylogenetic trees, and provides in computer programs each phylogenetic tree can be presented in linear form using a series of nested parentheses enclosing the names of nodes, separated by commas. Despite being the oldest of the three, the Newick format is one of the most common standards for recording phylogenetic trees. Used by applications GARLI [Zwickl 2006], PAUP [Swofford and Begle 1993], PHYLIP [Felsenstein, 1989], PROTML [Adachi and Hasegawa 1992], TREE-PUZZLE [Schmidt et al. 2002], MEGA [Tamura et al. 2013], iTOL [Letunic and Bork 2006] and others.

Nexus format [Maddison et al. 1997] is modular and its purpose is to allow upgrade and expansion. The file structure is made up of individual blocks - eg. TAXA contains information about the names of the taxa, CHARACTER contain sequences or morphological data; MATRIX - matching; TREE - the tree itself, written as a Newick string. The format is used by a number of software applications for phylogenetic analysis, including MacClade [Maddison and Maddison, 1997], PAUP [Swofford and Begle 1993], COMPONENT 2 [Page, 1993], SplitsTree [Huson and Wetzel, 1994] and Genetic Data Analysis [Lewis and Zaykin, 1996].

PhyloXML [Han and Zmasek, 2009] is specifically designed for storage and analysis of phylogenetic trees and associated information. Its main advantage is the ability to store except the structure of the tree and additional information about the nodes and edges of the tree. Using this standard data sharing is facilitated considerably and allows stored information to be displayed and processed with non-specialist to work with phylogenetic trees software tools. Unlike the Nexus, PhyloXML is structured as a hierarchical cluster of branches, each report corresponds to a node, and the group of all branches joining the root forms the whole tree. The format is used by applications Archaeopteryx [Han and Zmasek, 2009], HyperTree [Goodman and Sequin 1981], TreeGraph 2 [Stöver and Müller 2010], jsPhyloSVG [Smits and Ouverney 2010], Treevolution [Santamaría and Therón 2009)], iTOL [Letunic and Bork 2006], EvolView [Zhang et al. 2012].

3 Ensembl phylogenetic trees

Phylogenetic data in Ensembl are available in the form of gene phylogenetic trees downloadable as a single file in the so-called Ensembl multi format (EMF). It contains multiple phylogenetic trees, separated with two right slashes (//). The tree topology is written as a Newick string ("structural" panel) and is preceded by an "informative" panel containing information for each gene in the tree. The two panels are separated by a "DATA" line. The informative panel consists of space separated fields - beginning of the line (SEQ), name of the species (Species), name of untranslated sequence (Translation ID), location in the genome (Chromosome), beginning (Start) and end (Stop) of the gene, the direction of reading (Strand), name of the gene (gene ID).

4 Organising and editing the dataset

The phylogenetic data was processed with either custom-written Python 2.7 scripts (analyses based on supplementary data) or with the ETE Toolkit 2.2 (analyses based on the tree itself) which uses the same programming language [Huerta-Cepas et al., 2010]. The first step before the analysis involves indexing the trees by adding a consecutive number to DATA divider between information and structural. This marking every tree with a unique number allows their easy track to the division of the array of different groups in the next stages of the analysis.

The raw data was processed according to several criteria: 1) Trees containing genes from only one species were considered not informative and removed. 2) Genes that do not belong to plants were traced and removed from the trees. 3) Genes located in scaffolds and plastids were also removed. Exceptions were made only if a scaffold was listed in Ensembl as containing high percentage of genome sequencing data. After each change in the gene content of a tree it was first reconstructed and then a verification of the first criterion was carried out.

5 Homology between genes

After processing, the array search criteria, it is brought into a suitable form for carrying out the next steps of the analysis. The second stage - the extraction of information about evolutionary relationships between genes in trees, can also be seen as a preparatory because the information it used in tracking and predicting syntenic functions.

The task was used built-in function of ETE to predict ortho and paralogs. ETE has two such algorithms - Tree reconciliation [Page and Charleston, 1997] that relies on a comparison of the genetic tree with predefined types of wood and Species overlap [Huerta-Cepas et al. 2007], which uses overlapping Kladova and no need for tree species.

Given that data gene trees was chosen the second option. He performed the function get_descendant_evol_events, of the class PhyloNode. It detects all events of duplication or speciation occurring after a node (in this case after the root), and displays them for each tree as two lists of relationships one-to-one, one-to-many and many-to-many, respectively for orthologs and paralogs. Given the nature of the key genes are repeated many times in the resulting ortho and paralogous groups therefore need further processing of lists in order to reduce the groups to one-to-one, wherever possible.

The algorithm used to detect homologous relationships crawls all nodes of the tree in the direction from the root to the leaves, so that the same gene (leaf, end unit) is included in so homologous groups as the evolutionary events (intermediate nodes) which it away from the root of the tree.

In the direction from the root of the phylogenetic tree to the end node is increased and the degree of homology between the genes. Therefore the results with the greatest severity information appear groups of type 1-to-1, followed by 1-to-many, and with the lowest weight - many-to-many. They were separated from the remaining groups as follows - for each gene counting is done in all homologous groups where present, then one of them selects the one with the smallest total number of genes in it - in the ideal case 2. This was done for both groups - orthologs and paralogs.

Discovering homeology is part of a further processing and the first stage of it is searching the list of paralogs of relationship one-to-one and one-to-many containing only the genes of a species for which it is known from further literature, it is polyploid. The second condition is the differentiation of homeology paralogs from within the species, which is based on the acknowledgment that the genes in the group belong to different sub-genomes.

It should be noted that prior to selection of the closest homologous groups, all those involving animal genes were eliminated from the results. This does not affect in any way the accuracy, since trees are present animal genes, they are in any case separated in a stake which is the external group of stocks containing plant genes.

Establishing links between homologous genes is an intermediate stage of the overall study and, like the previous phase of collecting statistical information and “cleaning” of the array of trees, may be referred to the preparation of input data for the subsequent stages of analysis. As in all stages working with trees in digital form rather than visuals, where homologous relationships between genes would be clearly traceable, it is important to be able to be identified homologs of a gene at any one time. Regardless of preparatory role, predicting counterparts, especially the orthologs is essential for solving the main tasks, namely syntenic relationships between genes, predicting the functions and demand C4 genes.

Homologous relationships underlying syntenic approach to track the loss of duplicated genes. They are an essential source of information for predicting the functions of uncharacterized genes as information about their probable function is provided by precisely annotated their orthologs. The search for genes involved in C4 photosynthesis also depends on the relations between the homologous genes of the four species studied, as it is based on differences in the number of the replicate genes.

6 Conclusion

All operation concerning the evolutionary relationship between genes - looking for homologs predicting functions, tracking loss / gain of genes, detecting topological templates are based on the structural panel. Although not relevant to the structure of the tree, the informative panel was, however, essential for grouping and filtering the dataset, as well as mapping additional information from external sources (syntenic groups, functional annotations, expression data). The described methods and scripts are successfully implemented in pattern recognition for specific traits and functional annotation studies in plants.

Acknowledgments. The presented work has been funded by the Sofia University SRF within the “Methods for Automated Semantic Annotation of Heterogeneous Datasets” Project, Contract No. 15/2016.

References

1. Archie J, Day W H, Felsenstein J, Maddison W, Meacham C, Rohlf F J, Swofford D (1986). The Newick tree format. Online: http://evolution. genetics. washington. edu/phylip/newicktree. html.

2. Doyle J J, L E Flagel, A H Paterson, R A Rapp, D E Soltis, P S Soltis, J F Wendel, 2008. Evolutionary genetics of genome merging and doubling in plants. Annu Rev Genet. 42: 443–461.
3. Haas B J, A L Delcher, J R Wortman, S L Salzberg, 2004. DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics 20(18): 3643-6.
4. Huerta-Cepas J, J Dopazo, T Gabaldón, 2010. ETE: a python Environment for Tree Exploration. BMC Bioinformatics, 11:24.
5. Kersey P J, J Allen, M Christensen, P Davis, L J Falin, C Grabmueller, D Seth, T Hughes, J Humphrey, A Kerhornou, J Khobova, N Langridge, M McDowall, U Maheswari, G Maslen, M Nuhn, C K Ong, M Paulini, H Pedro, I Toneva, M A Tuli, B Walts, G Williams, D Wilson, K Youens-Clark, M K Monaco, J Stein, X Wei, D Ware, D M Bolser, K L Howe, E Kulesha, D Lawson, D M Staines, 2014. Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic acids research, 42 (D1): D546-D552.
6. Lyons E and M Freeling, 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. The Plant Journal 53:661-673.
7. Lyons E, B Pedersen, J Kane, M Alam, R Ming, H Tang, X Wang, J Bowers, A Paterson, D Lisch, 2008. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids. Plant Phys 148: 1772–1781.
8. Moreno-Hagelsieb G, V Treviño, E Pérez-Rueda, T F Smith, J Collado-Vides, 2001. Transcription unit conservation in the three domains of life: a perspective from Escherichia coli. Trends in Genetics 17 (4): 175–177.
9. Tamura K (1994) Model selection in the estimation of the number of nucleotide substitutions. Molecular Biology and Evolution 11:154-157

Information Extraction from Articles Related to Crime Events

Lyubka Genova

Sofia University „St Kliment Ohridski“,
Faculty of Mathematics and Informatics, Bulgaria
lyubka6ah@yahoo.com

Abstract. The paper presents a system which is developed to find named entities potentially connected with crime events. It collects articles describing crimes which happened around the globe. Lists of people, organizations and places are extracted and further ordered by relevance to the concept ‘crime’. The concept itself is semantically defined by a list with ‘golden terms’ - terms which have close meaning to it or are describing entities connected usually to crime events. To estimate the degrees of proximity of a page and of a phrase to the concept are used several ranking factors based on Cosine similarity and TF-IDF-score. The documents are ordered based on similarity estimation and their content is annotated. Named entities which are connected with the major event described in the document are discovered and their distribution over the whole set of documents is displayed in a graphical way.

Keywords: Information extraction, Conceptual description, Document ranking, Named Entities recognition, Cosine similarity, TF-IDF-score, Sentiment analysis

1 Introduction

The society is being tortured continuously by terrorist acts all over the globe. There is a huge need for the ability to foresee when, where and by whom mass murders or other violent actions will take place. Analyzing information about past events has a key role in finding regularities and connections between events and parties involved.

There are plenty of articles containing ‘Breaking news’ spread out via different news agencies and social media. The system described in this paper collects and evaluates such articles. People, organizations and locations are recognized, scored according to a criteria of closeness to the concept ‘crimes’ and collected. A dataset is built with the annotated content of the documents and lists of relevant named entities are created. Different correlations between the extracted data are presented graphically and a flexible GUI allows new searches and visualizations to be easily created.

In Section 1 are described the purpose of the project and the organization of the paper. Further in Section 2 are presented the tools which are used to implement the solution as well as the software architecture of the system and the communication between the different modules in brief. In Section 3 are described the preparatory stages of the process when the documents are collected, filtered and preprocessed to extract some initial lists of named entities potentially linked to the event. The main steps of the algorithm are listed in Section 4. Results presented visually may be seen in Section 5. Some possibilities for further development of the project are discussed in Section 6.

2 Project Specification

The main goal of the discussed project is to collect articles which describe actions of violence and to extract information from them for the location of the event, the people and organizations that are involved or took responsibility of the terrorist act.

2.1 Implementation

The project is implemented in Java and is integrated with several other systems which provide ready-to-go functionality for some of the sub-tasks. Those systems and the reasons why they were chosen are listed below.

Apache Nutch [1] is used as a web-crawler to collect and parse the web-pages. The reasons for this choice of a crawler are that it is stable, could be easily configured and allows integration with HBase, MongoDB and CouchDB.

GORA [2] is used as a connection layer between Nutch and MongoDB

MongoDB [3] is chosen to persist the data as it can be used to store large volumes of data which has little structure just as the case with the web-pages is. It could be easily scaled up which will be needed in the further development of the project. The documents stored in it could be indexed and searches done quite fast. It is also easily integrated with ElasticSearch and the other components which are needed to make the whole system.

GATE [4] is used for text manipulation and annotation of the collected documents. It extracts named entities from the documents which are further used as candidates for the parties really connected with the central crime event described in the document.

ElasticSearch [5] is used to index the documents and provide fast searches within their content and features.

Kibana [6] is used for visualization of the results of the findings and to make part of the searches within them more comfortable for the client.

JAVA [7] is used to integrate all the modules and to perform all the custom manipulations on the data.

2.2 Architecture

A software system was created to accomplish the goals of the discussed project. It consists of few modules each of which is responsible for the implementation of a different stage of the processing of the content of a web page (**Fig. 1)**. Having this segmentation from architectural point of view helps to reduce the complexity and leads to clear separation of responsibilities.

The core of the system is the Processing module which coordinates the other modules and passes the partially-processed content from stage to stage. The Nutch module collects the web pages and performs initial parsing. The Core module takes its output and gives it to the GATE sub-system. GATE outputs lists of extracted named entities which together with the content itself are analyzed and ordered by relevance by the Similarity module. The DB module is responsible for the saving, updating and reading of the documents in the MongoDB. After all the processing is done, the rated documents are inserted into an index using Elastic search. Visualizations of the data are presented by the tool Kibana.

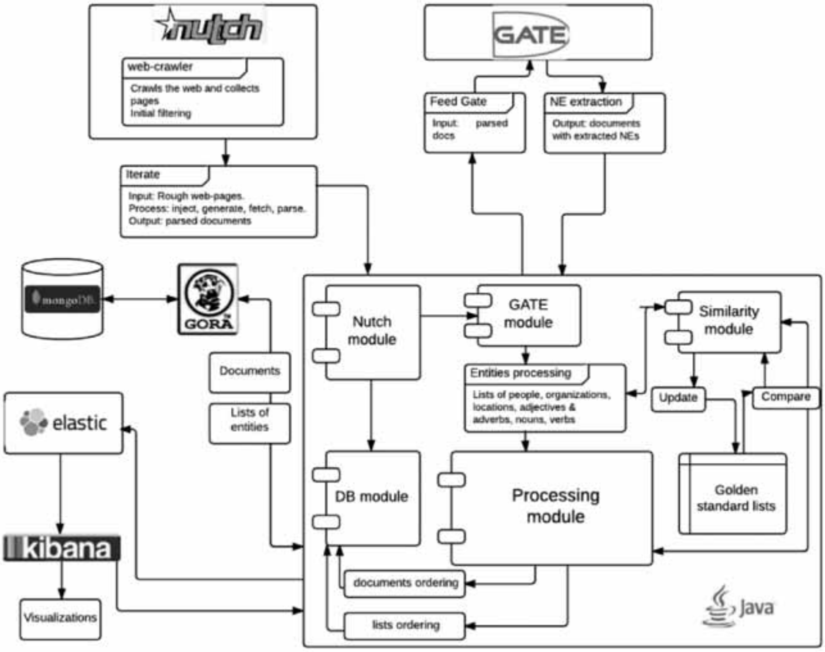


Fig. 1 *System architecture and coordination between separate modules with their own responsibilities*

3 Algorithm

The algorithm implemented by the system consists of several main stages [78, 89, 910]. Few URLs are fed into a seed file; the content of the web pages is collected, cleaned and parsed by the web-crawler. The document content is evaluated and if it passes a threshold of relevance then the document is saved. Named entities are recognized, lists of key phrases are collected. The extracted terms are evaluated, ordered and restricted to produce the final datasets created by the system. The evaluations are based on similarity to a 'Gold standard' set of terms which is continuously being enriched with new highly scored terms from top documents. The similarity scores are based on Cosine similarity and TF-IDF-score of the content and title of the document as well as the lists of terms collected for each document. Further details are given below.

3.1 Crawl with Nutch

The role of Nutch here is to find and parse into plain text the content of pages with breaking news. It does that via a pipeline of jobs to process URLs and page content. The following steps are iterated several times:

- Injector job – takes as input URLs from a seed file
- Generator job – makes a list with the links that will be downloaded as content
- Fetcher job – for each link of the above URLs fetches the content and the metadata of the web page
- Parser job – cleans up the content of the pages from all HTML tags and presents it as a plain text
- DB updater job – saves the data produced on the previous step into the database

3.2 GATE processing

Text processing of the plain text content of the downloaded pages is done with GATE (General Architecture for Text Engineering). The purpose of it is creating initial lists with named entities (people, locations, organizations) that could be related to the central crime event of the page.

After every five iterations of the steps of the Nutch module, processing with the GATE pipeline is performed once:

- Document Reset PR
- Annie English Tokenizer
- Annie Gazetteer
- Annie POS Tagger

- Annie NE Transducer
- Annie Orthomatcher
- Annie Pronominal Coreference

As a result are created the lists with the named entities and two additional helper lists – one with adjectives and adverbs and one with nouns and verbs that are considered relevant. They are later used as components of the score used to calculate the relevance of the document to the topic.

4. Main Processing

In this part of the paper is described the work of the main processing module of the system. It coordinates all the other modules and performs the key stages of the information extraction. The major steps that it performs are: evaluation of the document relevance, scoring entities, filtering and ordering of the lists of relevant terms, enriching the 'Gold standard' lists, creating the final datasets used by the output module.

4.1 Input

As input are taken:

- The documents (each with its content in plain text and the lists with named entities and helper lists of terms) provided by GATE,
- Three special documents – they are created initially manually and enriched with content automatically further. Those are: the so-called 'Golden standard' document containing terms tightly connected with the topic 'crime event'; the 'positive' list containing different parts of the speech with positive semantics; the 'negative' list containing different parts of the speech with negative semantics [11]. All terms in the documents here are presented by the roots of the words and the documents are presented as vectors of words. All the similarities are found in the vector space defined in this way.

4.2 Processing

For each document is found the Cosine similarity of its vector and the Golden standard vector [12, 13].

All documents which have Cosine similarity <= 0 are filtered out. The meaningful documents are sorted in descending order and passed to the next stage of processing. With each of them are associated lists with key phrases that were previously provided by GATE. Those lists are five and contain people, locations, organizations, adjectives and adverbs, nouns and verbs. They are analyzed, ordered and limited based on relevance of each phrase in them.

Phrase ranking

For each phrase in any of the five lists an estimation of closeness to the topic 'crimes' is generated.

First a cosine similarity is found by:

$$\begin{gathered}\text{total\_cosine\_similarity\_of\_phrase} = \\ 2 * \text{cosine\_sim}(\text{ vector (root\_of\_the\_phrase), vector (golden\_standard) }) \\ + \text{ cosine\_sim}(\text{ vector (root\_of\_the\_phrase), vector(negative\_terms) }) \\ - \text{ cosine\_sim}(\text{ vector (root\_of\_the\_phrase), vector(positive\_terms) })\end{gathered} \quad (1)$$

Then another measure of proximity is calculated – the TF-IDF-score towards the whole collection of documents [142, 154], where:

TF is:

$$\text{TF} = 0.6 + 0.4 * \text{TFd} / \text{maxTFd} \quad (2)*$$

* TF is the normalized value of the frequency of the phrase term in the document divided by the maximum value of the frequency of any term in the document. Normalization is used to neutralize the anomaly which appears if the document is too long and has many repetitions of the same phrase. The smoothing factor used is 0.6 and its role is to control the weight of the (TFd / maxTFd).

IDF is:

$$\text{IDF} = \text{total\_number\_documents} / \text{number\_documents\_containing\_the\_phrase} \quad (3)$$

And finally:

$$\text{TF-IDF\_of\_phrase} = \text{TF} * \text{IDF} \quad (4)$$

Now a total estimation of proximity is found for each key phrase of the lists for each relevant document:

$$\text{total\_score\_of\_phrase} = \text{total\_cosine\_sim\_of\_phrase} + \text{TF-IDF\_of\_phrase} \quad (5)$$

The lists with phrases are ordered decreasingly by this total estimation and limited to 10 elements.

Enriching the Golden Standard

Initially the Golden Standard document is filled manually with terms tightly connected to the concept 'crime'.

The information about how close each phrase in each document is relevant to the topic is utilized to enrich the Golden standard list. For each document in top 100 automatically are collected top 2 phrases of the People, Organizations, and Locations lists and top 3 phrases from the Adjectives-Adverbs and Nouns-Verbs lists. The roots of the words containing the term are put into the Golden Standard list if not already there. Manually are removed some less relevant ones. (This step will be further removed, but currently is needed as the Golden Standard has a key role and no compromises can be made with it).

Final phase

The work of the GATE module and the Similarity module are consequently repeated but on the next iterations the estimation of relevance is made according to the new list of golden terms and the result is not only an improved golden terms list but also annotated content of the documents which is serialized and saved into the database.

4.3 Result of the processing

Several ranking features of each document are found and saved:

- Cosine similarity of the vector containing the words in the title,
- Cosine similarity of the vector containing the words in the content,
- Total similarity for each collection from the 5 GATE lists,
- Total estimation of co-relation between the positive/negative lists and the Adjective-Adverbs/Nouns-Verbs lists.

The estimations are normalized within the range 1-100 and a final relevance estimation of each document is calculated as:

$$\boxed{\begin{array}{c}\text{total\_relevance\_of\_document} = \\ 2 * \text{cosine\_similarity\_of\_content} + \text{cosine\_similarity\_of\_title} \\ + \text{total\_score\_of\_adjectives\_adverbs} + \text{total\_score\_of\_nouns\_verbs}\end{array}} \quad \textbf{(6)}$$

5 Results

A collection of documents ordered by relevance to the concept ‘crime’ is created. Their content is annotated – people, organizations and locations listed in them and close to the crime are recognized. New terms describing the concept ‘crime’ are extracted and added to the concept description.

5.1. Elastic index with documents

The documents are stored in MongoDB and are indexed with ElasticSearch. The lists with key phrases are de-serialized to fields from that index. The fields could be divided in two main groups – one which shows the terms that were extracted and the other – which stores the scores of the separate terms or the whole list of terms.

Following this informal division there are columns such as the *GateAnnotations_Token_AdjAdv* column which contains the Adjectives and Adverbs connected with the crime event and the *GateAnnotations_Token_AdjAdv_Total* containing the scores of the terms.

```
{„nutch“:{„aliases“:{},
„mappings“:{„doc“:{„properties“:{„ContentAnnotated“:{„type“:“string“},
„CosineSim_Content“:{„type“:“float“},
„CosineSim_Title“:{„type“:“float“},
„GateAnnotations_Address“:{„type“:“string“},
„GateAnnotations_Address_Total“:{„type“:“ float „},
„GateAnnotations_Date“:{„type“:“string“},
„GateAnnotations_Date_Total“:{„type“:“ float „},
„GateAnnotations_Location“:{„type“:“string“},
„GateAnnotations_Location_Total“:{„type“:“ float „},
„GateAnnotations_Organization“:{„type“:“string“},
„GateAnnotations_Organization_Total“:{„type“:“ float „},
„GateAnnotations_Person“:{„type“:“string“},
„GateAnnotations_Person_Total“:{„type“:“float“},
„GateAnnotations_TokenAdjAdv“:{„type“:“string“},
„GateAnnotations_TokenAdjAdv_Total“:{„type“:“float“},
„GateAnnotations_Token_AdjAdv“:{„type“:“string“},
„GateAnnotations_Token_AdjAdv_Total“:{„type“:“float“},
„GateAnnotations_Token_VN“:{„type“:“string“},
„GateAnnotations_Token_VN_Total“:{„type“:“ float „},
„Total“:{„type“:“float“},
„Total_PosNeg_AdjAdv“:{„type“:“float“},
„Total_PosNeg_VN“:{„type“:“float“},
„anchor“:{„type“:“string“},
„boost“:{„type“:“string“},
„cache“:{„type“:“string“},
„content“:{„type“:“string“},
„digest“:{„type“:“string“},
„host“:{„type“:“string“},
„id“:{„type“:“long“},
„title“:{„type“:“string“},
„tstamp“:{„type“:“date“,
„format“:“strict_date_optional_time||epoch_millis“},
„url“:{„type“:“string“}}}},
„settings“:{„index“:{„creation_date“:“1467544885387“,“number_of_shards“:“5“,“number_of_replicas“:“1“,
„uuid“:“wuZGcWeFRd-sx55QGS2GHw“,“version“:{„created“:“2020099“}}},“warmers“:{}}}
```

5.2. Kibana

Different visualizations of the data are created and grouped in several dashboards. They represent groups of specific views over the collected and annotated texts and help the user to derive more easily conclusions about the data at a higher conceptual level.

A table with the main view for the collection is shown in **Fig. 2.** The documents are sorted in descending order by the total relevance rank, and the different intermediate metrics are displayed as columns. The user can see also a pie chart with the distribution of the total score across the whole set of documents.

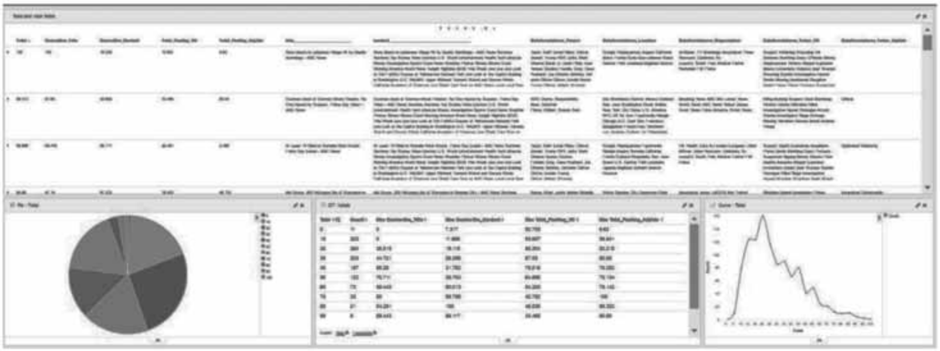


Fig. 2 *Intermediate and Total scores for the title, content and the lists of terms*

If we want to see the most often used Nouns - Verbs and Adjectives - Adverbs in the highest ranked documents we could take a look at **Fig. 3.** The table shows as rows the Nouns - Verbs and as columns the Adjectives - Adverbs sorted in descending order of number of documents they are used in and the intensity of the color of the cell is defined by the average value of the final score.

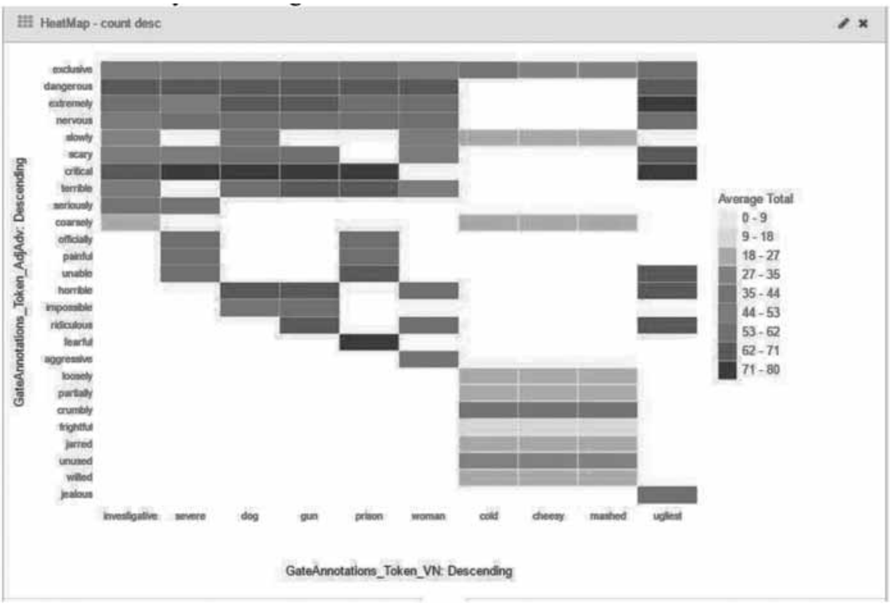


Fig. 3 *Heat maps of dense terms in high-ranked documents*

Representation of the collections of Adjectives - Adverbs and Nouns - Verbs is shown in **Fig. 4** and **Fig. 5**. In each of them is displayed information about the distribution of the terms in the two lists over the whole collection. In the columns is shown as follows: in the first column the phrases are displayed as tags with different size depending on the number of documents they are present in; in the second column is shown the co-relation between the intermediate scores for the different phrase groups; in the third column is shown the histogram of the total score broken into sections for each word according to the highest rank of the total score.

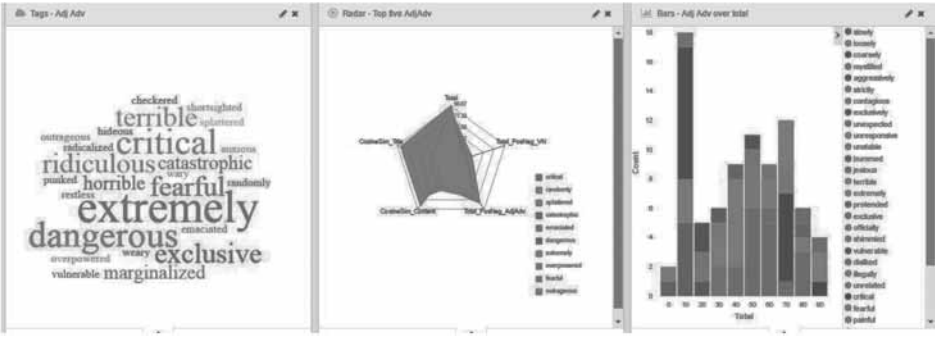


Fig. 4 *Distributions and co-relations of Adjectives – Adverbs*

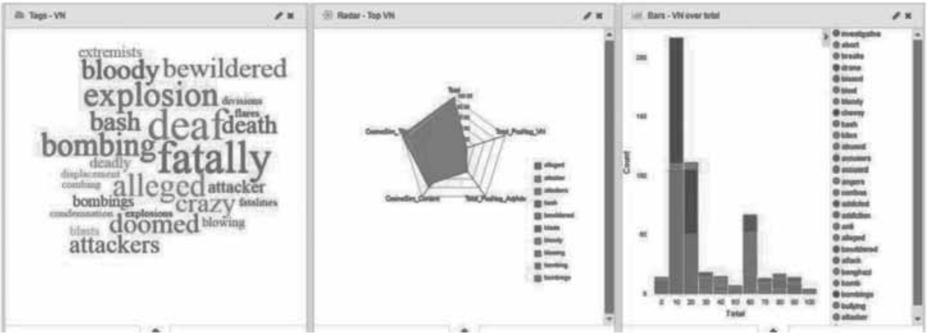
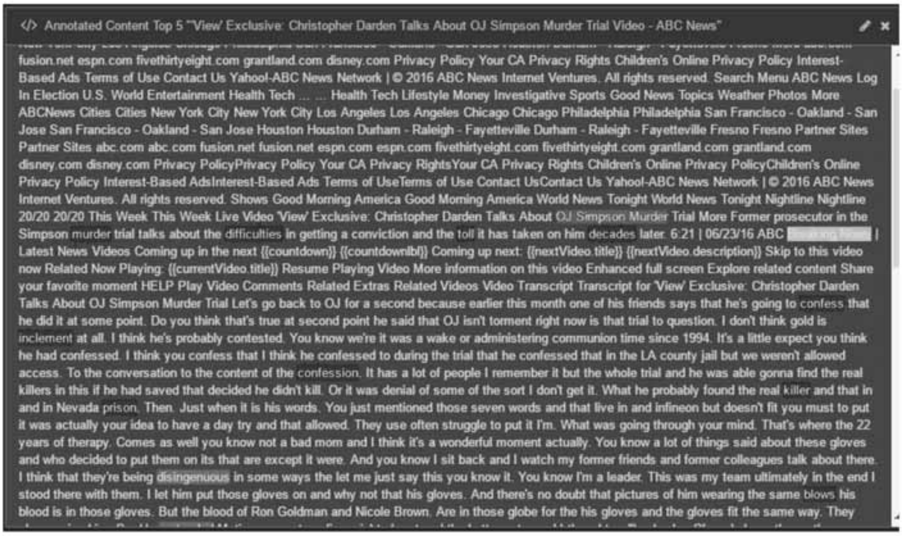


Fig. 5 *Distributions and co-relations of Verbs - Nouns*

Last but not least comes the view with the annotated content of the documents. The different types of extracted entities which are connected with the crime event are highlighted in different color. For the view are chosen the following documents: one from top 5 with a total estimation value of 80, one with total = 60, one with total = 40, one with total = 20 and the last one is with a total estimation value of 4. For brevity in **Fig. 6** is displayed only part of the content of the page ranked in top 5.


</> Annotated Content Top 5 "'View' Exclusive: Christopher Darden Talks About OJ Simpson Murder Trial Video - ABC News"

fusion.net espn.com fivethirtyeight.com grantland.com disney.com Privacy Policy Your CA Privacy Rights Children's Online Privacy Policy Interest-Based Ads Terms of Use Contact Us Yahoo!-ABC News Network | © 2016 ABC News Internet Ventures. All rights reserved. Search Menu ABC News Log In Election U.S. World Entertainment Health Tech ... ... Health Tech Lifestyle Money Investigative Sports Good News Topics Weather Photos More ABCNews Cities Cities New York City New York City Los Angeles Los Angeles Chicago Chicago Philadelphia Philadelphia San Francisco - Oakland - San Jose San Francisco - Oakland - San Jose Houston Houston Durham - Raleigh - Fayetteville Durham - Raleigh - Fayetteville Fresno Fresno Partner Sites Partner Sites abc.com abc.com fusion.net fusion.net espn.com espn.com fivethirtyeight.com fivethirtyeight.com grantland.com grantland.com disney.com disney.com Privacy PolicyPrivacy Policy Your CA Privacy RightsYour CA Privacy Rights Children's Online Privacy PolicyChildren's Online Privacy Policy Interest-Based AdsInterest-Based Ads Terms of UseTerms of Use Contact UsContact Us Yahoo!-ABC News Network | © 2016 ABC News Internet Ventures. All rights reserved. Shows Good Morning America Good Morning America World News Tonight World News Tonight Nightline Nightline 20/20 20/20 This Week This Week Live Video 'View' Exclusive: Christopher Darden Talks About OJ Simpson Murder Trial More Former prosecutor in the Simpson murder trial talks about the difficulties in getting a conviction and the toll it has taken on him decades later. 6:21 | 06/23/16 ABC Breaking News | Latest News Videos Coming up in the next {{countdown}} {{countdownlbl}} Coming up next: {{nextVideo.title}} {{nextVideo.description}} Skip to this video now Related Now Playing: {{currentVideo.title}} Resume Playing Video More information on this video Enhanced full screen Explore related content Share your favorite moment HELP Play Video Comments Related Extras Related Videos Video Transcript Transcript for 'View' Exclusive: Christopher Darden Talks About OJ Simpson Murder Trial Let's go back to OJ for a second because earlier this month one of his friends says that he's going to confess that he did it at some point. Do you think that's true at second point he said that OJ isn't torment right now is that trial to question. I don't think gold is inclement at all. I think he's probably contested. You know we're it was a wake or administering communion time since 1994. It's a little expect you think he had confessed. I think you confess that I think he confessed to during the trial that he confessed that in the LA county jail but we weren't allowed access. To the conversation to the content of the confession. It has a lot of people I remember it but the whole trial and he was able gonna find the real killers in this if he had saved that decided he didn't kill. Or it was denial of some of the sort I don't get it. What he probably found the real killer and that in and in Nevada prison. Then. Just when it is his words. You just mentioned those seven words and that live in and infineon but doesn't fit you must to put it was actually your idea to have a day try and that allowed. They use often struggle to put it I'm. What was going through your mind. That's where the 22 years of therapy. Comes as well you know not a bad mom and I think it's a wonderful moment actually. You know a lot of things said about these gloves and who decided to put them on its that are except it were. And you know I sit back and I watch my former friends and former colleagues talk about there. I think that they're being disingenuous in some ways the let me just say this you know it. You know I'm a leader. This was my team ultimately in the end I stood there with them. I let him put those gloves on and why not that his gloves. And there's no doubt that pictures of him wearing the same blows his blood is in those gloves. But the blood of Ron Goldman and Nicole Brown. Are in those globe for the his gloves and the gloves fit the same way. They

Fig. 6 *Annotated content of a document ranked in top 5 of all documents in the collection*

6 Further development

The project is an initial step to building a system for finding potentially dangerous people on the internet and their involvement into organizations planning or executing terrorist acts. There are many aspects that will be further improved: the quality of recognition of Named entities and their relation to the crime event; the classification of the acts of violence per degrees of impact, analysis of the location where the event took place to be and creating temporal maps of such events; collecting comments from people in the social networks over the events with biggest impact (as number of victims); analysis of the graph of friends of the people whose comments show distinguished sentiment either in a negative or a positive way, etc.

References

1. Nutch documentation, http://nutch.apache.org/ (visited on February 24, 2017)
2. Gora documentation, http://gora.apache.org/current/gora-mongodb.html (visited on February 24, 2017)
3. MongoDB documentation, https://docs.mongodb.com/ (visited on February 24, 2017)
4. GATE documentation, https://gate.ac.uk/ (visited on February 24, 2017)
5. ElasticSearch documentation, https://www.elastic.co/guide/index.html (visited on February 24, 2017)
6. Kibana documentation, https://www.elastic.co/products/kibana (visited on February 24, 2017)
7. Eclipse documentation, https://eclipse.org/
8. Search Engine with Apache Nutch, MongoDB and Elasticsearch, http://www.aossama.com/ (visited on February 24, 2017)
9. Building a Java application with Apache Nutch and Solr, http://cmusphinx.sourceforge.net/2012/06/building-a-java-application-with-apache-nutch-and-solr/ (visited on February 24, 2017)
10. Nutch as a Web data mining platform, http://www.slideshare.net/abial/nutch-as-a-web-data-mining-platform (visited on February 24, 2017)
11. Bo Pang and Lillian Lee: Opinion mining and sentiment analysis Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1–135
12. Machine Learning: Cosine Similarity for Vector Space Models (Part III), http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/ (visited on February 24, 2017)
13. Simone Teufel Natural Language and Information Processing (NLIP) Group, Term Weighting and the Vector Space Model
14. Rajendra Kumar Roul, Omanwar Rohit Devanand, S. K. Sahay Web Document Clustering and Ranking using Tf-Idf based Apriori Approach
15. Manning, Raghavan, Schutze: An introduction to IR. ISBN 0521865719 (2008)

A SAT Approach for Task Assignment

Vesela Angelova

Institute of Mathematics and Informatics – Bulgarian Academy of Sciences,
Acad. G.Bonchev Str., Bl. 8, 1113 Sofia, Bulgaria
vaa@math.bas.bg

Abstract. The paper shows the need for compliance in the management of the human resources with conflicting requirements, such as skills and restrictions or an insufficient number of people. A specific approach has been chosen and its programming model, developed for practical purposes, has been described. The work with the model will enable managers to choose an appropriate team and it will be also used to maintain in the work process for the task assignment. The model is oriented to the wishes of the developers and can be integrated into more extensive task management software.

Keywords: Software Technologies, Task assignment, Agile Methodology, SAT solver.

1 Introduction

Software production is related to practical situations in which several potentially conflicting constraints must be satisfied. Several recent software development methodologies such as Agile require that initially or weekly, tasks are assigned to potential teams. In many of these assignments, the team consists of scattered people in a company with limited number and resources. In such situations, external help or at least an advice is needed to form teams and assign tasks. A main problem in solving these assignment tasks is the numerous requirements that need to be satisfied in the team selection process. Furthermore, in the work process when the number of tasks is more than the number of people, there is a need to rearrange the teams depending on the progress of the project. Thus, the team assignment process may need to be performed continuously during a project lifecycle. These changes may stem from the software team (internal causes, e.g., failure) or context (external events, e.g., increasing requests from users). Such a system is required to monitor itself and its context and to recommend how to relocate assignments. In practice, such team assignment is carried out periodically (for example, every week) by the team itself and its manager or Scrum Master for every new subtask.

A generalized task assignment problem has been studied in a number of previous applications and was found to be an NP-hard equivalent to integer linear programming. This problem is typically solved with "branch and bound"

V. Dimitrov, V. Georgiev (Editors): ISGT'2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

methods such as [10], but these techniques are generally considered to be slow in the presence of hard to satisfy constraints [11] even with applying more recent techniques [12]. The main reason is that they attempt to solve a more general problem where the main goal is to optimize the cost as opposed to satisfying the constraints.

In this work, we are interested in task assignment, which concerns human resource (assigning software development tasks to programmers) and rapid change processes supposing agile methodologies for programming. We take into account various constraints imposed by people and encode the problem using only Boolean variables.

This means that the value of each variable ranges in TRUE/FALSE (also denoted shortly as 0/1). We express the constraints on the variables by propositional logic formulas on the variables (with AND, OR and NOT). Given a valuation of the variables and a formula, we search a satisfying assignment of the variables such that the formula evaluates to TRUE (if such an assignment exists). This means that we search those variables which satisfy the constraints. SAT is short for "satisfiability". At this time, we are not going into details on complexity and NP-hardness of the problem, because this problem is solved by the SAT solving programs and is guided by many heuristics and competitive rules (Section 2). Thus, our work can benefit from state-of-the-art techniques to speed up SAT solving.

Our paper is using SAT solvers for recommending people for a developing team (Section 3) and for assigning tasks to the selected team (Section 4).

2 About SAT

The SAT problem uses a set of clauses built from a propositional language with n variables and gives an assignment of these variables that satisfy all clauses.

A formula is given in conjunctive normal form (CNF), which is a conjunction of disjunctions of literals, where a literal is a variable or its complement. Each disjunction of literals is called a clause. For example, the following is a formula on three variables with two clauses:

$$F(x1,x2, x3) \equiv (\neg x1 \vee x3) \wedge (x2 \vee x3). \quad \textbf{(1)}$$

A formula is said to be satisfiable if it can be made TRUE by assigning appropriate logical values (i.e. TRUE, FALSE) to its variables.

SAT is one of the first problems, that was proven to be NP-complete. Despite the fact that no algorithms are known that solve SAT in polynomial time for all possible input instances, many practical problems can actually be solved rather efficiently using SAT-solvers [3]. In practice, SAT solvers first simplify the given formulas in order to reduce the number of variable assignments that need to be checked with a brute-force exponential algorithm. Additionally, modern SAT solvers include a number of optimization techniques such as local search, random search, genetic algorithms and other heuristics.

Despite its theoretical hardness, strong drivers such as an open SAT competition led to the success of SAT solvers much to the surprise to many in the computer science community [4]. Today we can find SAT technology in many areas like formal methods, artificial intelligence and games [9], bioinformatics, design, security. In this work, we propose to use SAT to efficiently assign tasks in the context of a software team. In our implementation, we use the sat4j [7] solver that was implemented in Java and enables easy integration with our Java code. Using SAT solver enables us not only to provide one task assignment, but to enumerate all possible assignments. To do this, we perform multiple SAT queries where in each query by explicitly disallow all solutions found so far by adding a clause to the CNF formula.

The main contribution of this work is a practical implementation of the SAT-approach and support for specific procedures in software development.

3 A Choice for Developing Team: Skill Constraints

Consider a company with a personnel of *n* software developers X*1,* X*2,…,* X*n*. We will visualize each developer by a column in a table. Each line of a table will list a particular skill or restriction of a developer. Typically, the skills of developers are known in advance and the restrictions are set by the project owner. These skills and restrictions form clauses that are passed to the solver. Here is the essence of choice: the more restrictions are imposed, the fewer choices remain. For explanation these limits are not just filters which cut a group of people, they are nonlinear functions for preferences that should be modeled in the language of solvers.

When developing a new project, its unique requirements lead to unique needs for people with diverse skills. On the other hand, other project restrictions such as costs, quality and speed impose constraints that make developer resources transient [5].

Let Table 1 lists the constraints of developers with respect to their ability to solve a particular task. For example, we have a description of the programmers by few skills (design, testing, and management) and one special feature - the inability to collaborate. If we consider only the skills, this task is combinatorial and there are several solutions. But the SAT solver allows combining people with few skills or restrictions.

From such a table, we introduce one Boolean variable for a developer. Thus there are in total four Boolean variables. Each row of the table encodes a SAT clause. The last row denotes a constraint that developer X1 cannot collaborate with developer X3. The resulting formula is:

$$F(X1, X2, X3, X4) \equiv (X3) \wedge (X1 \vee X2) \wedge (X1 \vee X4) \wedge (\neg X1 \vee \neg X3)\ . \qquad (2)$$

Table 1. Constraints for people

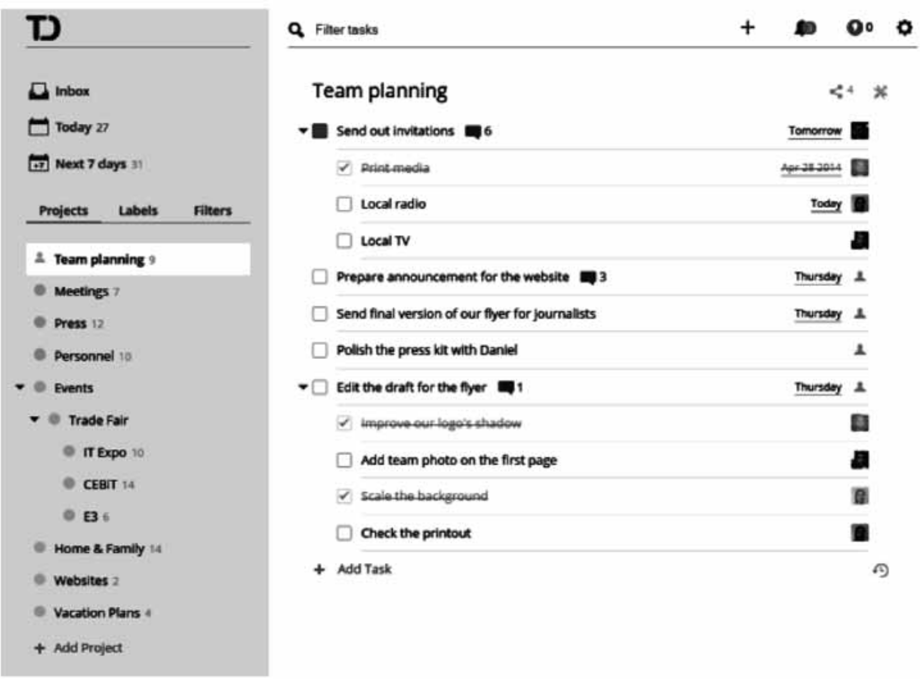
	X1	X2	X3	X4
Skill 1	0	0	1	0
Skill 2	1	1	0	0
Skill 3	1	0	0	1
Collaborate	-1	0	-1	0

This example has only one assignment of developers for the task: {X2, X3, X4}.

Of course, the number of constraints can be increased by increasing the required skills and assessments of each for everyone, as well as the gender proportion. Additionally, we can have a restriction for particular or minimal number of the team members. There are many advisers for finding the right team, for example experience and even mistakes made by person are a great source of insight and start-up [2]. It is good policy to take into account personal preferences, formal requirements, desires and psychological habits as constraints. In a future version we plan to add the ability to assess the developer not only with Boolean variables as well as allow developers to make assessments of the skills of their peers.

4 Assigning the Tasks

This strategy is inspired by Steve Jobs' deeply ingrained concept of a Direct Responsible Individual. Having a "DRI" is a simple, yet highly effective way to ensure accountability for task execution and to alleviate communication issues by clearly showing who is in charge of what [1].



Team planning

- ■ Send out invitations 6 — Tomorrow
  - ☑ ~~Print media~~ — ~~Apr 28 2014~~
  - □ Local radio — Today
  - □ Local TV
- □ Prepare announcement for the website 3 — Thursday
- □ Send final version of our flyer for journalists — Thursday
- □ Polish the press kit with Daniel
- □ Edit the draft for the flyer 1 — Thursday
  - ☑ ~~Improve our logo's shadow~~
  - □ Add team photo on the first page
  - ☑ ~~Scale the background~~
  - □ Check the printout

Fig. 1. Delegating sub-tasks to fellow team members [1]

A similar approach we can see in Scrum methodology with Backlog refinement for Sprint meetings [8]. Scrum differs from "DRI" by self-organization. We expand this table by the proposals from everyone which tasks (a few) he/she could take. So, there will be an opportunity for a better distribution of tasks within the team.

So far, we looked at assigning developers to a single task. However, a project includes a number of tasks that need to be assigned and worked on in parallel. In Table 2 we give one such example with three developers and three tasks. The value of each cells of this table denotes if a developer has a personal preference to be assigned to a particular task. These decisions are often made weekly in a Scrum team.

This table is encoded to a Boolean formula in a slightly different way. Each cell here is a Boolean variable. We denote these variables with x11, x12, x13, x21, x22, x23, x31, x32, x33. Here the first index is for the task, the second index is for the programmer. Next, we want each task to be assigned to one person. This means that if one of the two variables in a particular row is TRUE, the other is FALSE or both cannot be TRUE. There are two constraints (clauses) for each row with two '1' in order to choose only one. The clauses for the first row and for the third row are:

$$(x11 \lor x13) \land (\neg x11 \lor \neg x13) \quad \textbf{(3)}$$

$$(x31 \lor x32) \land (\neg x31 \lor \neg x32). \quad \textbf{(4)}$$

For second row the clause consists of one variable x23.

The similar constraints we extract for each column. The resulting formula combines all constraints for the rows and for the columns.t

Table 2. Preferences of 3 developers

	X1	**X2**	**X3**
Task 1	1	0	1
Task 2	0	0	1
Task 3	1	1	0

There is one solution for this example – {x11, x23, x32}. This means that Task 1 is assigned to X1, Task 2 to X3 and Task 3 to X2. This solution satisfies all personal preferences and can be accepted.

When the number of tasks and people are equal, this problem is matching problem [6], which can be solved in polynomial time. The SAT solver however solves a more difficult task that allows combining the constraints from Section 3 and the assignment constraints. Furthermore, encoding the problem as SAT allows combining more tasks than people or more people than tasks as well as returning all possible solutions to a problem.

The next example (Table 3) is searching a solution for three tasks and four programmers. If every programmer needs to participate, we will remove the restrictions for rows with two '1', so each of two variables can be TRUE.

Table 3. Preferences of 4 developers

	X1	**X2**	**X3**	**X4**
Task 1	1	0	1	0
Task 2	0	1	0	1
Task 3	1	0	1	1

There are five solutions for the task assignment: {x13,x22,x31,x34}, {x11,x13,x22,x34}, {x11,x22,x33,x34}, {x13,x22,x24,x31} and {x11,x22,x24,x32}.

Note that when combined with a constraint that programmers X2 and X4 are incompatible to work together (encoded as the clauses (¬xi2 ∨ ¬xi4) for i=1,2,3), the number of solutions reduces to only the first three.

Table 4. Preferences of 5 developers

	X1	X2	X3	X4	X5
Task 1	1	1	1	0	0
Task 2	0	0	0	1	0
Task 3	0	1	1	0	1

Finally, consider the example on Table 4 with three tasks and five programmers. This example without additional constraints includes four solutions: {x11,x13,x24,x32,x35}, {x11,x12,x24,x33,x35}, {x11,x12,x13,x24,x35} and {x11,x24,x32,x33,x35}. If we don't allow combining 3 programmers for a task (encoded as the clauses (¬xi1 ∨ ¬xi2 ∨ ¬xi3) for i=1,2,3), this additional constraint reduces the number of solutions to the first two.

Such restrictions can be placed in the case of a smaller number of programmers compared to the number of tasks. Furthermore, if all tasks should be assigned, we can remove the restrictions for more than one '1' in a column. For example, it is not allowed to delegate three sub-tasks to one programmer or, two specific sub-tasks not to be delegated to one programmer.

The SAT approach gives opportunity for flexibility and feedback. It is necessary to refine sub-tasks if there is no solution. In the case of more solutions, the team decides which one to choose according to personal, group or external considerations. This corresponds with self-regulation in Scrum or allocating the resources by the project manager.

5 Conclusions and Future Work

In this work, we propose an automated SAT technique for assignment of tasks to teams. In practice, software development teams are not larger than 10-12 developers and tens of tasks. This leads to task assignment problems consisting of up to a few hundred Boolean variables, which is well within the capabilities of model SAT solvers such as sat4j that we use [4].

The main contributions of this work are solving a problem in agile software developing and encoding the problem and the solutions by the terms of SAT approach. We believe that unlike existing task management process through its life cycle, our approach complies with the wishes of the developers and thus should lead to better teamwork. Our practical application of SAT technology in the software development process (especially agile), serves as an advisor for team choice and later for regular tasks assignment.

In the future, we expect to extend the technique beyond Boolean formulas. More difficult constraints such as linear equations can be handled by other solvers such as Satisfiability Modulo Theory (SMT). These extensions may allow us to express variables in other logic (for example, evaluation of people) or temporal properties of the system (for example the development of people).

Of course, the SAT-approach is suitable for other areas with teamwork.

References

[1] The Todoist Team, May 20, 2014 https://blog.todoist.com/2014/05/20/using-steve-jobs-strategy-to-assign-tasks-in-todoist/

[2] Elvekrog J., June 11,2014, Entrepreneur http://www.entrepreneur.com/article/234648

[3] Le Berre D. Understanding and using SAT solvers. http://resources.mpi-inf.mpg.de/departments/rg1/conferences/vtsa09/slides/leberre1.pdf

[4] Malik Sh. and L.Zhang. Boolean Satisfiability: From Theoretical Hardness to Practical Success. Communications of the ACM. August 2009 (Vol. 52 No. 8), pp.76-82

[5] Mundra A. et al., Practical Scrum-Scrum Team: Way to Produce Successful and Quality Software. In Proc. of 13th Int. Conf. on Computational Science and Its Applications, IEEE, 2013, p.119-123

[6] Dan Ma, February 18, 2010, https://probabilityandstats.wordpress.com/2010/02/18/the-matching-problem/

[7] SAT4j. The Boolean satisfaction and optimization library in Java

[8] Schwaber K., M.Beedle, Agile Software Development with Scrum, ISBN 0-130-67634-9, 2001

[9] Weber T. A SAT-based Sudoku Solver, 2005, https://www.lri.fr/~conchon/mpri/weber.pdf

[10] Ross and Soland, “A branch and bound algorithm for the generalized assignment problem,” Mathematical Programming, vol. 8, no. 1, 1975, 91-103

[11] 11.Ramon E. Moore, Global optimization to prescribed accuracy, Computers and Mathematics with Applications, (Vol. 21, 6–7), 1991, pp.25-39

[12] Katsutoshi Hirayama, A New Approach to Distributed Task Assignment using Lagrangian Decomposition and Distributed Constraint Satisfaction. Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA.

System SlopeStabBG for Calculation of Stability Factor in Opencast Mines

Mariana Trifonova

University of Mining and Geology “St. Ivan Rilski”, Sofia, Bulgaria

Abstract. The present article introduces original Bulgarian software product SlopeStabBG aimed at calculation of stability factor of opencast mines’ boards. This product offers opportunities for examination of stability without restrictions on the geometry of mine’s profile. The results are obtained using strength criterion of Mohr-Coulomb for various variants of sliding surface including excursive, generated by user on the basis of expert evaluation, defined by the concrete situation like system of cracks, weathering and so on. Probabilistic character of strength characteristics of the massif is taken into consideration as well through integration of Monte-Carlo Method. The product works in the environment of Autodesk products.

Keywords: stability factor, strength characteristics, opencast mines, software product SlopeStabBG.

1 Introduction

The problem with determination of stability of the slopes and the boards at opencast mines is particularly topical. There are various approaches for resolving of this problem most common of which are the following:

1. Application of finite elements method. This method offers good and relatively reliable results but is not always applicable especially for mines where clay is prevailing. For instance for the conditions of mines “Maritza-Iztok” up to the moment there’s no adequate decision.
2. Use of empirical formulas based on strength criteria of Hoek-Brown or Leon-Torre and geotechnical GSI classification [1], [2].
3. Calculation of stability factor on the basis of strength criterion of Mohr-Coulomb with assigning of sliding surface’s type [3].

Of course, other decisions can be found in the literature, too, for example, rock massif modeling as a set of rock blocks and their analyzing through “n-angles functions method” [4].

Combination of two of aforesaid methods is possible, too [5].

Accent of the present article is presentation of developed by the author system SlopeStabBG, applying the third approach and its comparison with popular similar systems.

2 Specialized Software for Evaluation of Stability of Opencast Mines Offered in the Market

The development of specialized software aimed at evaluation of stability of opencast mines', quarries' and natural slopes began in the 70^{ies}. More popular products, applying the third approach in this field, are SLOPE, SLIDE, PRIZMA and some others.

SLOPE can be used for calculation of stability factor of multilayered slopes (but not a board) and offers opportunity for rendering an account of seismic forces, external loads and hydrodynamic pressure. But the problem with the choice of surface of sliding is not well determined here. More precisely such a choice is not made. Four digits - r_1, r_2, r_3, r_4 are given only which represent limiting (minimum and maximum) values for the abscissa of start and end points of the landslide. The program allows representation of the sliding surface as an arc of circle only.

Although the program Slope facilitates considerably the calculations related to slopes' stability evaluation it features the following disadvantages:

- lithological types' geometry is completely entered in text mode, which complicates the work and is a prerequisite for mistakes when the amount of data is bigger;
- the program works for a slope only, not for a board;
- the evaluation of slopes' stability is made through determined values for body's characteristics;
- there's no choice of calculation model;
- there's no information about number and nature of examined surfaces.

The program SLIDE is composed of three modules: MODEL (creation of slope model), COMPUTE (calculation) and INTERPRET (interpretation of the results obtained).

All three modules can be started as independent programs, but they exchange between each other information on file principle.

The program offers opportunity to render an account of: level of underground water, evenly distributed and concentrated loads. User can as well choose the sliding surface among some defined in advance. Excluding the lack of opportunity to apply a probabilistic approach for calculations, the rest of disadvantages of SLIDE are absent in Slope. Profile data can be entered both as text in MODEL environment and through exchange drawing file (*.dxf), created in advance in AutoCAD but the strict limitations on this file's structure impose difficulties for the user.

SLIDE is actually a professional product featuring rich functionality and can be applied to many cases. As its disadvantages can be mentioned the unsatisfactory flexible opportunities for input of vector graphic information. Furthermore the product offers no opportunity for probabilistic evaluation of stability and respectively of landslide risk.

The program PRIZMA is developed by a team of University of Mining and Geology, lead by Prof. P. Zlatanov and Assoc. Prof. G. Trapov in the end of last century in DOS environment and, of course, it works in console mode. The program offers opportunity to calculate the stability factor for:

- single layer slope on 6 types of sliding surfaces;

- multilayer slope on 2 types of sliding surfaces when the layers are horizontal;

In both variants the program offers opportunity for construction of probabilistic model of the slope.

3 Description of System SlopeStabBG Calculating Stability Factor of a Board

This product is developed by the author and actually is an aggregate of additional commands in AutoCAD or other Autodesk's products (without Light versions) supporting object LWPOLYLINE. Commands are grouped in a separate menu and toolbar "OTKOS". It is developed in Visual Lisp and C# using DCL (Dialog Control Language) for the Lisp's windows.

The product doesn't impose limits on the number and geometry of lithological types. As sliding surface can be used both curves of second and third degree (arc of circle and parts of parabolas) and any excursive polyline generated by the user. This is particularly useful at cracked rocks where the sliding surface has no regular geometric form but follows the system cracks. Actually these two aspects: the excursive geometry of the profile and the opportunity for excursive form of the sliding surface are the main advantages of the product.

To apply the functions of a module, contour of the board and the borders of lithological types should be drawn first through standard AutoCAD commands. If the sliding surface is defined by the user its contour should be drawn. All aforesaid contours could be represented by excursive lines. The only requirement is that they must be objects of LWPOLYLINE type, i.e. lightweight polylines in the point of view of AutoCAD. The user has not to close them, but their contours must be drawn in the way allowing closing of each of the zones of the lithological types programmatically. Special organization according to the layers is not needed i.e. each polyline can be in any layer. The contour of each lithological type is not necessary to be one closed polyline as needed in the program SLIDE, since this is a prerequisite for errors, especially when the lithological types are adjacent. Moreover the information is doubled. Since an opportunity for probabilistic approach is envisaged, if necessary data are available they must be entered in advance in text files. Upon these preliminary operations the functions of "OTKOS" menu can be put in use, namely:

S *SetMat* – a command for calculation of strength parameters of each of the lithological types. It operates like most of commercial products for strength simulations. The input is done in a dialog box and the compulsory fields are as follows:

- Volume weight, angle of internal friction and cohesion. Opportunities are envisaged both for determinate input in relevant text field and probabilistic whereat the system offers the user to choose the file containing accumulated for the parameter in question set of values and finds its cumulative function;
- Pseudonym (up to 10 symbols often under the form of abbreviation). It is used afterwards in the command Lith.

The command should be started at least as many times as many are the lithological types. If needed, the input information can be saved in an external file. Respectively, if in a previous session some information about a given lithological type has been entered, this information can be taken from a file.

Lith – through input of an internal point generates a closed contour as a border of a lithological type and fixes a "material" for it. In principle the two commands *SetMat* and *Lith* could be united in one but the variant of two commands is chosen because the availability of more than one zone of a given lithological type is possible and thus a repeated dialogue of entering of strength parameters can be avoided.

Proper work of the command requires as follows:

- Objects of LWPOLYLINE type to be drawn on the screen. They should enclose the area of the relevant lithological type;
- Information on strength parameters of each of the lithological types should be entered through the command *SetMat*;

StepEdge – offers the user to select the contour of the board which must be drawn in advance.

SlideSurf – a command for choosing of sliding surface.

Calculate – basic command of the product. It doesn't require additional input from the user. The ratio between retaining and sliding forces in the massif is calculated on the basis of information entered through above mentioned commands. If some of strength characteristics are defined probabilistically then sliding risk is calculated, too. Calculations are based on Mohr-Coulomb criterion and are done according to formula (1), after values of all variables in it have been *automatically* calculated in advance:

$$\eta = \frac{\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{m_i} S_j \gamma_j tg\varphi_j\right)\cos\alpha_i\right) + \sum_{j=1}^{k} c_j l_j}{\sum_{i=1}^{n} T_i} \qquad \textbf{(1)}$$

where:

- n – number of the lamellas that the area limited by the contour of the board and sliding surface is divided into;
- k – number of parts that sliding surface is divided into by the lithological types;
- m_i – number of parts that i–lamella is divided into by the lithological types;
- $\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{m_i} S_j \gamma_j tg\varphi_j\right)\cos\alpha_i\right)$ – sum of friction forces;
- $\sum_{j=1}^{k} c_j l_j$ – sum of cohesion forces;

- $\sum_{i=1}^{n} T_i = \sum_{i=1}^{n} P_i \sin \alpha_i$ – sum of tangential forces;
- P_i – weight of the lamella;
- α_i – angle of the slope of the tangent to sliding surface in the middle of the lamella basis;
- S_j – area of the portion of lamella situated in a given lithological type;
- l_j – length of the line of sliding of a given type.

The choice of program environment is largely determined by the following considerations:

- input of graphic information in text mode (for instance by coordinates from keyboard) is very uncomfortable and is a requisite for errors;
- availability of means for editing of entered vector graphic information is desirable, which assumes application of some CAD system;
- in Bulgarian mines AutoCAD or some other product of the firm Autodesk is largely used, which means that automatic import of a part of necessary input information like slope and/or board contour, borders between different lithological types and to some extend specific user's sliding surface is possible.

Currently the developed software product is presented under a form easy for use without seeking a commercial visualization.

4 Comparison Between the Developed Software SlopeStabBG and Some of the Programs Offered in the Market

The calculations for the stability factor done using programs Slope, Slide 5.0 and the developed by the author product show commensurable values. The conducted experiments show that the difference in the values obtained using Slide 5.0 and SlopeStabBG varies between 5,4% and 11,7% for homogeneous slope and respectively between 4,4% and 17,9% for non homogeneous slope. Taking into consideration that values for the stability factor obtained by the developed product are lower this can be considered an advantage from security viewpoint.

SlopeStabBG gives results almost identical to those of program PRIZMA in the cases where PRIZMA can be applied.

5 Conclusion

Most of above listed disadvantages of program products offered in the market are taken into consideration in the course of the present software development which gives the latter the following advantages:

1. Opportunity for excursive number and geometry of the lithological types in the massif;
2. Opportunity for evaluation of board (slope) stability at various forms of the potential sliding surface, including excursive form according to the expert evaluation of the user and taking into consideration the system cracks, weathering and others.
3. Opportunities to enter not only determined values of input data but also for taking into consideration of probabilistic character of one, a number of values or all values taking part in the calculations.
4. Easy manipulated by the user. Complicated geometric problems are resolved through elementary dialogue with the user.

The system SlopeStabBG can be used both in the process of operation of opencast mines and for research works and training of students in the University of Mining and Geology.

References

1. Hoek E., Estimating Mohr-Coulomb Friction and Cohesion Values from the Hoek-Brown Failure Criterion, Journal of Rock Mechanics and Mining Science, V. 27, No 3 (1990)
2. Hoek E., Strength of Rock and Rock Masses, Journal of Rock Mechanics and Mining Science, V. 2, No 2 (1992)
3. Газиев Е., Механика скальных пород в строительстве, Москва (1973)
4. Kandov L., N. Iontcheva, M. Draganova, Clastic Mechanics Problems Related to the Analyses and Study of Boards and Slopes, 5-th National Open Pit Mining Conference with International Participation, Varna (1998)
5. Trifonova M., P. Roussev, Calculation of Slope Stability According to Fissenko Methed on the Bases Of Hoek&Brown and Leon&Torre Fairure Criterion and Geotechnical GSI Classification of Bieniawski&Barton, 5-th National Open Pit Mining Conference with International Participation, Varna (1998)

Software Engineering Management Education – Bringing Industry Standards to the University Program

George Sharkov[1] and Maya Stoeva[2]

[1] Faculty of Mathematics and Informatics, University of Plovdiv "Paisii Hilendarski"
24 Tzar Asen str., Plovdiv 4000, Bulgaria, and
European Software Institute – Center Eastern Europe, Sofia, Bulgaria, gesha@esicenter.bg

[2] Faculty of Mathematics and Informatics, University of Plovdiv "Paisii Hilendarski"
24 Tzar Asen str., Plovdiv 4000, Bulgaria, may_vast@yahoo.com

Abstract. Nowadays, especially in the field of Information Technologies (IT), to be a good specialist means being able to find the intersection point between theory and practice. Therefore, university programs in software engineering and development face the challenge of balancing the theory related to current and future industry demands with the sufficient practices to build the required competences. This paper presents our approach and experience in introducing industrial software development management models like CMMI (Capability Maturity Model Integration) [1] to university programs. The good practices are illustrated through real software projects at Faculty of Mathematics and Informatics of Plovdiv University. We discuss why quality and "QA" should be more than testing software and code and has to address the quality of processes through the complete development lifecycle. We outline which process areas of CMMI are suitable for cultivating basic software development discipline and how to exercise them in real life related projects. Specific role playing and teaming are demonstrated along with the concept of how proper motivation could go beyond the students' grades.

Keywords: CMMI, software engineer competences, software quality management, industrial standards, process improvement, software requirements, project management, requirement development, modern training methods

1 Introduction

One of the main challenges of the modern academic education in software engineering and informatics is bridging theory with practice and better prepare the students for real software industry dynamics of work, while yet keeping the interest and motivation for deeper understanding of theoretical and generic aspects. It is not anymore sufficient just to "transform" the knowledge into practical technical skills by introducing intensive exercises and cultivating technically savvy specialists.

Understanding the software product development lifecycle, related processes and roles, specific aspects of teamwork, as well as the variety of organizational models with related advantages and implementation burdens, are those expected competences that differentiate developers and software engineers from what students themselves address

V. Dimitrov, V. Georgiev (Editors): ISGT'2016. ISSN 1314-4855
Proceedings of the 10th Jubilee International Conference on
Information Systems and Grid Technologies, Sofia, 30. Sep. – 1. Oct., 2016.

as “coders”. Those gaps and insufficiencies of the ICT university education content are frequently underlined by the major ICT stakeholders in Bulgaria and the region. Similarly, a strong interest is demonstrated in the introduction of internationally recognized master programs and series of top-class ICT courses with particular focus on IT business, project and team management. Certain number of courses already address the problematic areas, but they lack synergy and unification, and are frequently up to the initiative (and availability) of the lecturers. The Master and BSc programs offered by the main Bulgarian universities are yet focused on rather technical than managerial profiles.

To address that gap and align with the global trends, a national wide project SEMP (“Software Engineering Management Program”) was launched in the year of 2010, coordinated by the European Software Institute – Center Eastern Europe (ESI CEE) and in collaboration with Carnegie Mellon University (Institute for Software Research and Software Engineering Institute), involving 5 leading Bulgarian universities [2]. Upon the initiation of the program, a survey conducted by ESI CEE under the Regional Competitiveness Initiative (RCI) project by USAID for Bulgaria and 7 countries in the region showed clearly underdeveloped training and knowledge in professional ICT educational areas: Project Management, Software Process Improvement, ICT Services Process Improvement, and Information Security (Fig. 1).

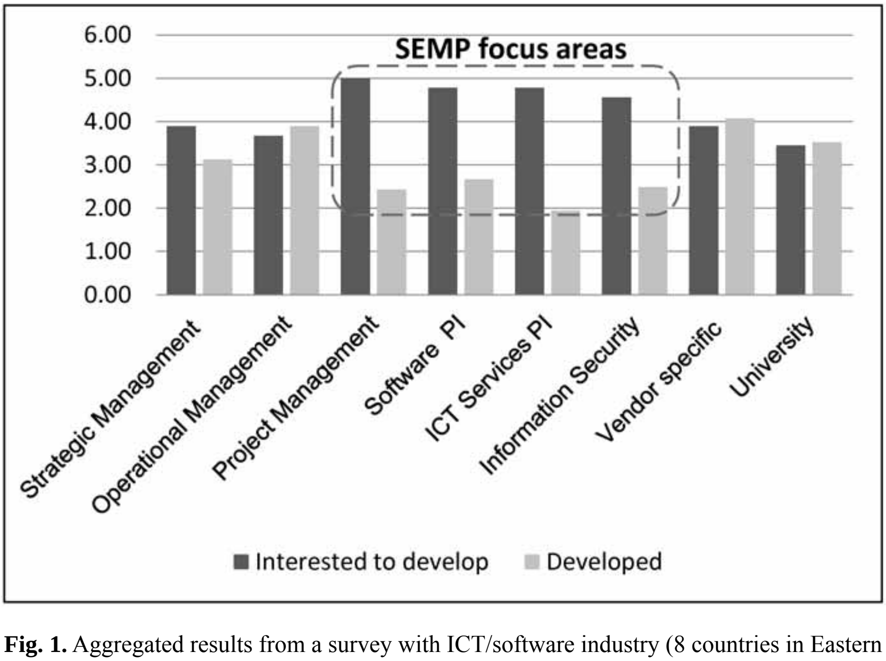


Fig. 1. Aggregated results from a survey with ICT/software industry (8 countries in Eastern Europe) on skills and competences expectations (bipolar scaling method, measuring negative to positive attitudes to a statement, *with values from 0 to 6, where 6 is “high”*), compared to the opinion on the current status for the period 2009-2013 (*“PI”* stands for Process Improvement, and *“University”* is for academic/theoretical foundation). SEMP is Software Engineering Management Program [2].

2 Background

During the pilot phases of SEMP (the years 2010 – 2016), more than 18 academic courses were introduced or built upon already existing courses to align to the standards and methodologies obtained through the know-how transfer from the Carnegie Mellon University (Institute for Software Research - ISR, Software Engineering Institute – SEI, others). Seven courses were implemented during the first phase (2012-2013), while others were gradually implemented after 2014 (as an elective or regular/core courses). An interuniversity and an industry-recognized certificates from the SEMP program were issued to all students who successfully completed the courses, supplementary to their scores and credits received from the university. In total since the beginning of the project more than 1000 students have completed SEMP courses. Several courses offered also an intensive (executive) format option, and were attended both by students and industry professionals, thus giving additional benefit from mixed teaming.

The approach chosen by the Faculty of Mathematics and Informatics at Plovdiv University (FMI-PU), also partner in SEMP, was one of a gradual implementation. In addition to the continual upgrading of existing courses to bridge the theory-practice gap and meet the new trends, like [3] and [4], a new course was first introduced for several years as an introductory elective course under the title "Software Process Quality Management". The course content was based on the materials and experience of similar courses under the SEMP program, delivered by ESI CEE, or implemented at Carnegie Mellon University (CMU), Sofia University (FMI-SU) and other SEMP partners [5], [6], [7]. The feedback, however, showed that students with mainly technical expertise cannot easily absorb the abstract process definitions and link the process areas and practices (as in CMMI model) with a real software project lifecycle and activities. Additionally, being an elective, the course admittedly was unable to provide the targeted common understanding and unified project management "language" and link with the other disciplines studied. Thus, the beneficial opportunity to unite practical tasks from other disciplines was not utilized at all. They remain formal as a "homework", exercising some individual skills but yet too far from the real industrial projects. Nevertheless, the course was considered as a significant asset for students when seeking employment in the IT sector (see feedback results in Fig. 2 at the end).

In order to align better the course with the development strategy of the Faculty and to address the above-mentioned shortcomings, a new "Software Quality Assurance" course was introduced as a core course of the 4th year of study in the BSc program of Informatics. Despite the usual association of "QA" with "testing", the course focuses on the quality of processes as a prerequisite for the quality of the product and how the process improvement leads to excellence. This course is also based on the CMMI for Development (CMMI-DEV) model, benefitting from its completeness and recognition by the industry as a de-facto standard, but further adapting the relatively "heavy" model to the knowledge and skills level of undergraduate students. The model was applied as a generic sample model, to which other modern methods and techniques like Scrum, "Disciplined Agile" with CMMI, Six Sigma, Kanban are consequently addressed in the course. The course also introduces business aspects such as “cost of quality” along with the need to balance process improvement and product quality with a brief introduction of Kaplan and Norton’s Balanced Scorecards, applied to IT and software industry.

For two years of delivery, the core undergraduate course “Software Quality Assurance “ introduces more than 200 students to processes improvement as a main factor for the quality of software products. The goal, however, is not to study the CMMI model. The model is used as a reference framework to describe the main processes along the lifecycle of a typical software project and cultivate software development discipline by exercising in life-related realistic projects. Such team-based realistic projects approach is typical for masters’ programs, such as the practicum or “studio project” at CMU-ISR [8], [9], or the MSc programs at New Bulgarian University [10] and FMI-SU [6]. Several effective “competition” style ideas and practices with industry involvement from similar SEMP-related courses at FMI-SU were adopted as well [6].

3 Motivation and Approach

Undergraduate students in informatics are typically interested in acquiring a wider range of practical competences and obtaining good grades, so they can jump quicker to industry positions. Naturally, they are more interested in attending lectures and courses that tackle more practical, rather than theoretical aspects. Most of them have certain practical experience working in companies or ad-hoc projects, which do not always represent good examples of project organization. Software development technologies and platforms rapidly evolve and require respective drastic revisions of academic content in order to remain “modern”. On the other hand, the logic behind software projects lifecycle is quite generic and the real challenge is how to “translate” in academic environment the abstract processes into practical implementation rules.

After collecting sufficient experience from the elective course “Process Quality Management” and similar courses of the SEMP program, and also from software industry (see also Fig. 1), we have proposed for the core course „Software Quality Assurance (Q.A.)” the following approach:

First: Focus on the *quality of processes* – introduction to a generic industry-oriented reference model, such as CMMI, as a key factor for software product quality (applicable to IT services as well). Students will have a good idea and a sense of the meaning of processes and company maturity, institutionalization, goals and practices, as well as generic approach and practices for organizational process definition and management.

Second: Dive deeper in selected process areas related to *software projects management*. Show students the logic and interconnectivity of project phases with the respective process areas and specific practices, with good practical examples of their implementation and “what if not” discussions.

Third: Studio development project – exercise the theory or abstract examples in *realistic collective projects*, organize students to work in teams and think about developing innovative applications on the basis of well-prepared documentation, compliant with another internationally approved standard like ISO/IEC/IEEE 29148:2011 [11]. Assign roles and responsibilities within processes (role playing).

Fifth: Challenge students' *creativity* – if the supply of industry-related project proposals and ideas are insufficient (typical for such an initial course stage), invent fictional software projects that focus on real business problems. Challenge established business models (kind of "hackaton"). Provoke and cultivate the *sense of a startup*.

Sixth: Encourage *self-organization* and management of teams, deciding on ideas, scope of projects, plans and deliverables. Provide mentorship (organizational and technical, involve industry and professionals from other areas) [3].

An important factor that helped us to incorporate successfully such a practical studio project was the approval of the course as a core course for BSc students in Informatics. This allowed stable team composition and result oriented collective impetus (turned to be impossible with elective format). Such approach is widely used in master's programs ([8], [9], [10]), where project and team management are in focus, and also reinforced by earning credits of critical importance for graduation.

However, since the major source of developers for the industry comes from undergraduate programs, the teaming and project culture is already a critical asset, as opposed to just an advantage. So it was beneficial for students to introduce it earlier, in undergrad program (see feedback in Fig. 2). Another motivation was that most of the students from the last year in university already have working experience in IT companies and have observations on "bad practices" as well. They are prepared to extend the understanding of quality assurance beyond the quality code producing and testing, and the importance of the established processes, quality of requirements, planning, monitoring and control for successful project delivery in time.

4 Course Goals, Focus and Expected Outcomes

The main goal of the „Software Quality Assurance (Q.A.)" course is to provide students in Informatics with contemporary theoretical and practical knowledge for developing quality software code based on well-defined and institutionalized processes, using CMMI-DEV as a reference framework model.

Our focus is to bring the real in-company working atmosphere to the university classroom by applying modern teaching methods, training techniques and style of organization in order to break the patterns of the ordinary teach-exercise-grade education system. We address the involvement of the different stakeholders - students, professors, future employers, government, end users, IT companies and clients working together based on the established "common language" – the language of quality.

We anticipate that proper positioning of the course within the BSc program in Informatics at FMI-PU and bridging it to the real-life industry demand, will better prepare the students to create and contribute to the success of their projects. Through managed processes and furthermore good software specifications, students are enabled to develop quality code for software services or products in time, while meeting client expectations and quality criteria.

Strategically, the studio project organization, combined with a better synchronization with other (more technical or theoretical) disciplines could absorb the usual individual practical tasks into studio project product aligned tasks.

5 Theoretical Background and Scope

The sound IT/software fundamental and technical-related background for the 4th year students at the Informatics undergraduate program is essential for the successful introduction of the „Software Quality Assurance (Q.A.)" course. The experience from earlier versions of the course taught as elective (SEMP at FMI-SU) shows insufficient preparedness of students from initial phases. We rely on their good knowledge and sufficient practical experience of working with different programming languages and software development environments, algorithms, design patterns, basic design and architecture principles, object-oriented models and tools, databases, etc.

The course introduces modern models and standards for process quality management in the field of software engineering and IT services. The major structure and content of the course is based on the CMMI-DEV model coming originally from Carnegie Mellon University – Software Engineering Institute, and well acknowledge as industry de-facto standard [1]. Although considered as a heavy industry model for big organizations mainly or criticized for not being "modern" or agile enough (among the "myths" about CMMI), the model provides a very well structured and complete reference framework for internal IT/software companies organization and the pathway to industry maturity. Of course, it needs careful selection and gradual introduction, with sufficient practical examples and exercises.

Another challenge was also how to adapt such a professional and normally very unattractive formal matter (close to the scope of a heavy 3-day Intro to CMMI course of the CMMI Institute) and make it affordable and inspiring for not so professionally experienced students. Fortunately, the experience of the lecturers and the SEMP program partners (most of them with extensive industry experience) along with the feedback acquired through more than 8 years of teaching similar elective course in 2 Bulgarian universities, helped a lot to optimize the content structure, and introduce the model gradually and just-in-time to answer the needs of the ongoing studio team projects, running in parallel to the lectures.

The *introduction part of the course*, however, is not technical or IT focused. On the contrary, it elaborates around competitive strategies, applicable for (any) business development, combined with a very quick introduction of the Balanced Scorecards method of Kaplan and Norton [12], tailored to fit the specifics of the IT/software industry. Nonetheless, the importance of the *digital dependency* in the modern society, the new "digital ecosystems" are also discussed, along with the concept of why software quality is of vital importance. Such topics quickly break the technology biased mindset of students, and achieve the understanding why compliance to standards or adherence to models is not a goal ("to build the ideal company") but a mean to achieve excellence in business and customer satisfaction. A detailed analysis of *"cost of quality"* is performed, and at the end students discover themselves how useful the statistics from process improvement programs is in order to build the software business profitable.

With this background, and following the lifecycle of a typical software product development project, a quick overview of various quality models and standards is made, based on the paradigm that the *quality of the product depends largely on the quality of the internal processes*. The *cost of quality* understanding is correlated to the *cost of defect and correction.* Thus, the advantage of detecting defects in all project related products (not just the code itself, but the specs, the plan, and also the *established related processes*) as early as possible during the project lifecycle is concluded.

The maturity of the processes as a differentiator for mature companies brings a natural basis for deeper introduction to CMMI. For the sake of better clarity, we use staged representation with Maturity Levels (limiting the scope to Level 2 and 3).

Maturity Level 1 (named also *Performed*) is a perfect illustration of companies in a "survival" mode. Here we usually have an open discussion session with students that work on their observations over such a typical picture - processes and deliverables are unpredictable, poorly controlled, and the mode is reactive, so project delays and unsatisfied customers are common.

Essential part of the course content is dedicated to Maturity Level 2 (*Managed,* or also referred to as a r*epeatable*). All process areas are reviewed in detail as a stage dedicated to successful project management. Goals and practices are described to implement managed, institutionalized processes and why we need their continuous improvement.

In order to "dive" deeper into the process areas from level 2, for illustration we follow the classical "waterfall" model and project lifecycle, which brings better clarity on the subject of planning (with options and techniques for establishing estimations, critical path and optimization logic, etc.). However, since the process areas of CMMI are adapted already to agile development style, good examples of "disciplined agile" are covered. Practical examples of other "modern" organizational frameworks, like Agile and Disciplined Agile, Lean Kanban (or more general – Lean development), Scrum are shown and encouraged to apply in studio projects.

We make detailed presentation of six process areas from Maturity Level 2:

Requirements Management (REQM)
Project Planning (PP)
Project Monitoring and Control (PMC)
Process and Product Quality Assurance (PPQA)
Configuration Management (CM)
Measurement and Analysis (MA)

However, to jump-start their studio projects and group work, we first go over two of them - *Requirements Management (REQM)* and *Project Planning (PP)*. This is the necessary minimum to start transforming the business ideas into draft project specifications, work breakdown structure with respective estimates, project plans (with Gantt charts). Consequently, with the progress of the work the topic of what progress monitoring methods could be applied and why corrective measures are needed is discussed, thus introducing naturally the PMC process area with its respective specific practices. Afterwards, students understand through experience what a "significant deviation" from the project plan means, how to follow corrective measures and also why *Configuration Management (CM)* is more than a source control.

Although the main focus is laid onto the specific goals and practices (that define the scope of the process areas), the generic goal and the 10 generic practices are repeatedly illustrated, finally giving the logical justification and link specific practices to be implemented for managing the *quality of processes* under *PPQA*. Some basic indicators needed for process improvement are also underlined by the mentors, so students can recognize the value of the practices under *Measurement and Analysis (MA)* process area. The seventh process area from Maturity Level 2 *(Supplier Agreement Management, SAM)* is only briefed, as in several cases a need to delegate or "outsource" some studio project tasks to other teams was identified.

From Maturity Level 3 (*Defined*) we cover only a selected subset (mainly around the requirements development and validation, peer reviews under *Verification*) as the students need them for the studio projects. In fact, since the very first initiation stage of discussing the project ideas, the teams already need some guidance on eliciting and formalizing "customer" requirements. Some hints come logically with specific practices from *Requirements Development (RD)* process area. *Validation (VAL)* process area becomes critical in few cases of real industry or business-related project ideas. Some of them are about “patching” or improving real systems in use (like a family hotel ERP management, for example) – a challenging task even for mature companies.

From the higher Maturity Levels 4 and 5 (*Quantitatively Managed* and *Optimizing*) only the goals and business benefits are discussed (again referring to the *Cost of Quality* structure).

6 Modern Training Methods and Tools. Results

With the adoption of student-centric and project based approach to teaching software quality models we have improved the style of coaching and course organization, which made students address this course as a “different one”, or the “interesting one”.

By referring to industrial models and standards (vendor neutral) we give a realistic picture of what industrial environment is (or should be). The overview of existing international software development and IT services standards and models, their applicability in different industries and for different purposes, as well as cross-mappings overcome the natural students’ fear of such a “bureaucracy”.

Familiarizing with one of the de-facto industry standards (CMMI) and exercising it through team-based realistic project conveys in fact an “industrial” atmosphere to the university classroom. Each of the ten lectures is literally transiting into related practical exercises that are not artificial, but structured around the studio projects. In addition, the obstacles or difficulties, which students meet during their project lifecycle become topics for discussion at the lecture session to follow, which in conclusion makes the activities and content dynamic and responsive to student’s needs and real case problems. A repository of modern tools (mainly open source) and techniques with proper samples is maintained [13], [14].

6.1 Teamwork Motivation and Role Playing

The two main factors to motivate students for teaming and make them collectively accountable for studio projects are:

- Participating in a successful studio software project is the key to receive higher final course assessment (score);
- Only teamwork is acknowledged (teams up to 5-7). No individual projects.

One very important and usually neglected aspect while introducing the processes, goals and practices is related to respective roles and responsibilities (sometimes imprecisely referred to as “job profiles”). In our course those are not only outlined, but simply assigned and exercised by the team members in the ongoing studio projects. Thus, competences based on knowledge and skills are acquired by experience in somewhat “gamified” (or role playing) environment.

Some typical roles (like project manager, architect, designer and developer) are easy to assign at the time of initial team composition, although the understanding of their responsibilities evolve along the advancement of processes details and practical needs. Others come with the progress of their studio project and real tasks and activities distribution. Our main function as mentors is to align those profiles with respective process areas activities and responsibilities, and also make sure that members with more than one “hat” (needed for smaller teams) act appropriately. Another engagement is to make sure that we systematically develop the ICT professional competences based on the newest European standard e-CF (European e-Competence Framework) [15].

6.2 Teams Decide on Project Scope and Development Phases

Studio projects are defined based on elaboration of real life business idea – either coming from industrial partners (preferred case) or invented by students. In both cases those ideas are not mature or digested enough to start programming. On the other hand, we have already similar to real life projects limitations – restricted timeframe, resources, client responsiveness, tools and skills. However, the purpose of the studio project is still to exercise the project management related processes, but with the view of some final realistic product or service. The deliverables expected include various project related materials (like product or service specifications, work breakdown structure, tasks and estimates, plans, assignments, test plans, etc.). Therefore, under our mentorship, we let teams discuss, propose and justify their project scope, develop their optimistic and pessimistic scenarios, and phases with milestones.

Some teams came with a brand new ideas with a real intention to turn them to evolve to real startup business. Others focused on real life business cases (including some family or small business management, or extensions or patching existing systems). Few problems came from supportive companies. Another option observed was to build a business case around the implementation of students’ final thesis for the graduation. In all cases the freedom to decide and self-organize was balanced with the team responsibility to deliver which required certain discipline, careful progress monitoring and corrective decisions.

It is difficult for students without practical experience to decide on complexity of functionalities, system architectures or modules to be included in the project. And here we have the proper place for requirements management and project planning practices, estimation methods, design architectures and knowledge acquired in other disciplines. The studio project scope is considered not only in terms of the functionality but also in view of levels of project and product design, wireframing, pilot prototype or complete application development.

Clearly, mentoring those activities is different from classical lecturing. In many cases the mentors, other affiliated professionals (or even other teams) have to play the role of customers, users, investors or other relevant stakeholders while critically discussing projects scope and progress. Having a common repository of sample projects, typical project development lifecycle steps, previous projects documents selection is valuable [16]. Running projects and discussion draft documents are also shared as part of the open competition.

We have also identified promising scenarios to link with other courses, which integrate front-end web development, server modules and services (including mathematics and sophisticated algorithms), combined with dedicated mobile apps – all to contribute for the same business idea, managed as a project under our studio course.

6.3 Documenting and Presenting the Work

Creating standardized documentation is part of the professionalism and competences usually underdeveloped in academic trainings. Here we introduce again a typical representative industrial standard for software requirements specifications [11], but directly as a framework for their particular studio project development. Other typical documents are also requested and created (work break down structure, Gantt charts and project plan) compliant to CMMI guidelines and by using a variety of open source tools.

Documenting some main internal processes around project management is also exercised, again with the practical value and focus on the specific teams and projects. The purpose is not to "show" nice schemes, but to result in implementing and determine at least few gaps or difficulties during the course of their studio project.

At the end of the course, each project is presented by the team before all the class and external partners involved. A team (not leader only) effort is required and evaluated in 2 directions – the project results with the tasks completed, but also from organizational perspective the processes establishment and improvement opportunities identified. Those are the 2 final ingredients of project success evaluation.

6.4 Using Contemporary Tools and Techniques

The student centric approach also requires a variety of contemporary interactive tools and techniques that should be provided to students [13], [14] in order to facilitate the "learning by doing" assimilation of the theory and enable documenting their studio projects. The repository contains also major documents templates (currently based on Google documents), like standardized software requirements specifications, sample plans, typical design patterns, etc.

The repository helps creating work-like environment and develop work break down structure (WBS) and project plans (for example, Gantt charts are created by using Smartsheet online application) [14]. When project plans are ready, we exercise with students how to create easily their wireframing with Moqups and Uxpin. Practically, all studio projects so far include some web or front-end interface design. Since this is the most visible part of their project implementation, but on the other hand the technical part is not a direct objective of our course, we give them a complete freedom to the technology and tools for this development.

The repository of tools, techniques, and documents samples is open and helps the team work, and improves significantly the working environment, accumulating knowledge and answers to frequently asked questions from previous academic years.

6.5 Competition and Unique Certificates

In order to encourage innovations and creativity and establish realistic business arena and "competitive" environment, we provoke challenges to the teams in many areas – like, for example, about the novelty of their team names, project titles, internal organization, best customer profile or business idea.

We let the audience decide on the awards, and also provide additional "bonus" scores and unique certificates by category. Successfully completed project teams receive also a SEMP recognized sign. Those competitions and certificates turned to be a huge motivation factor for students and inspiration for innovative and "crazy" ideas.

For two academic years of running the "Quality Code Assurance" course as described, we had 17 projects successfully completed - 6 for the first year (with less than 50% of students involved) and 11 for the last (close to 100% students involved). This growth shows clearly their motivation to play the "realistic project game" and compete. The project ideas evolved as well – although online shopping or e-commerce systems were the most popular, we have a business to business very serious application presented, two game portals, three mobile applications. There is also a nice prototype for a project planning tool, which development continue after the completion of the course and is created for the purposes of the course itself.

The most severe competition, however, was the one around the original team names. Not surprisingly, some of the winners already think about startup business with, like "TEAMofPoxies" (one of the game portals), "DirtyMuddy" (business to business application for hotels' bookings), "NanoLab" (Fantasy mobile/ desktop game).

7 Course Feedback

The growing interest and motivation of students to engage proactively in the studio software projects within the course “Software Quality Assurance” shows the relevance of the content and the approach. Our students could map theory to practice and develop valuable skills like: creating, developing and managing software project specifications and plans following industrial de-facto standard (CMMI-DEV), understanding and applying mature software development processes, analyze them for improvements, focus on defects prevention, work in teams with roles and responsibilities, sense the real industrial atmosphere and style of work (including “bonus” system for performance and creativity).

A summary from feedback forms received from FMI-PU students participating the last 2 years of the course delivery as a core course with studio projects is shown in Fig. 2, in comparison to the aggregated feedback from similar CMMI-based course at FMI of Sofia University. The essential differences between those two course formats are that the latter is elective (also for BSc, but mixed diverse years of study and specialties), and the practical teamwork is limited to group study and presentations (e.g. on process area or implementation).

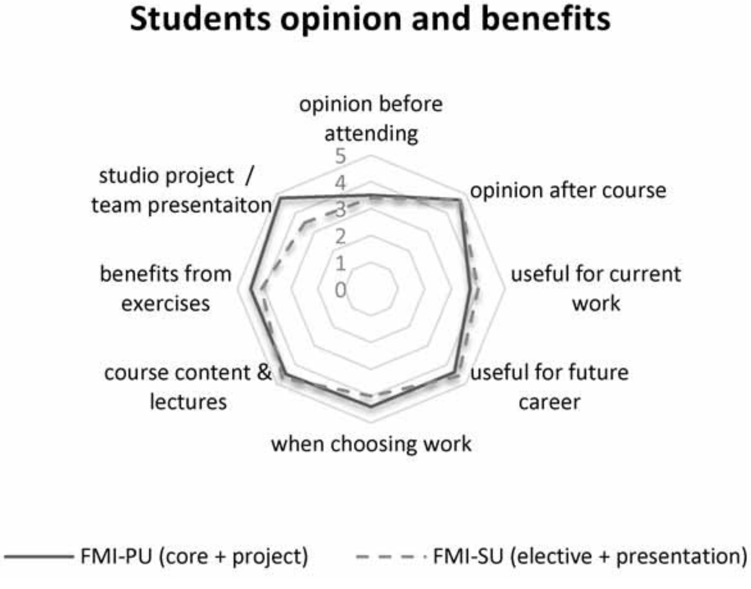


Fig. 2. Aggregated feedback results from software quality management courses at FMI, Plovdiv University (core course with studio project teamwork) and FMI, Sofia University (elective, with group study and presentations as teamwork).

In both groups the opinion on the course benefits and value for real work and employment are comparable (high). However, the significant difference in students' opinion on teamwork format clearly shows the advantage of studio project. We can also add from this group of students' open comments like: *"The teamwork and the practical knowledge gained on CMMI", "The knowledge I gained for my future projects", "The knowledge I gained, which I can apply in practice", "The teamwork and the simulation of an actual work environment".*

8 Conclusions

This paper presents our approach to better prepare students in software engineering and informatics in BSc university programs for real industry work on projects, and with deeper understanding of software quality and development lifecycle. An introduction of a core course „Software Quality Assurance (Q.A.)" for last (4th) year of study at the Faculty of Mathematics, University of Plovdiv is described, with the feedback from two years of delivery as compared to other similar courses and formats. The most appreciated advantages of the course consist of: (1) Introduction to industrial de-facto standard model (CMMI) and (2) Applying and exercising theory in realistic studio teamwork projects.

We have demonstrated how such a course could be used as a framework to align better theory with practice in other courses being taught, and unite "homework" individual exercises as building blocks of a team project related to real business area. The team work, although not really full time occupation, gave good enough experience about roles and responsibilities, as well as basic soft skills needed in industry.

References

1. CMMI for Development model - version 1.3, http://cmmiinstitute.com/resources/cmmi-development-version-13
2. SEMP (Software Engineering Management Program), http://semp.esicenter.bg
3. Rahnev, A., Stoeva, M., Arnaudova, V.: Educational Dynamic web site for nonspecialists. In: Synergetic and reflection in the mathematic education conference, Batchinovo (2010)
4. Stoeva, M., Krushkova, M.: Modeling and implementation of interactive web-based system for education and entertainment with games, From DeLC to VelSpace. In: International conference, FMI, Plovdiv University "Paisii Hilendarski", Plovdiv town, Bulgaria, ISBN: 0-9545660-2-526-28, p. 265-274 (2014)
5. Kaloyanova, K.: Design from Data: How To Use Requirements for Better IS Analysis and Design. In: Proceedings of the Int. Conference Informatics in Scientific Knowledge, pp. 189-197., (2012)
6. Kaloyanova, K.: Including Real Stakeholders at Students Projects. In: Proceedings of the 9th International Conference Computer Science and Education in Computer Science (CSECS), Fulda/Wurzburg, Germany, p. 55-59, (2013)
7. Sharkov, G., Asenova, P., Ivanova, V., Gueorguiev, I., Varbanov, P.: Evaluation of ICT Curricula using European e-Competence Framework. In: 10th Annual International Conference on

Computer Science and Education in Computer Science 2014 (CSECS), pp. 267-286, https://www.ceeol.com/search/article-detail?id=469775 (2014)

8. Lattanze, A.J.: Practice Based Studio. In: 2016 IEEE 29th International Conference on Software Engineering Education and Training (CSEET), (2016)
9. Root, D., Lopart, M., Taran, G.: Proposal Based Studio Projects: How to Avoid Producing "Cookie Cutter" Software Engineers. In: IEEE 21st Conference on Software Engineering Education and Training, Charleston, US (2008)
10. Ivanova, V., Leading a development team pilot for IT project management master's degree students, http://eprints.nbu.bg/1932/1/LDT_V.Ivanova.pdf
11. IEEE Recommended Practice for Software Requirements Specifications, Online ISBN: 978-0-7381-0448-5, http://ieeexplore.ieee.org/servlet/opac?punumber=5841 (2017)
12. Kaplan, R. S., Norton, D. P.: The Strategy-Focused Organization: How Balanced Scorecard Companies Thrive in the New Business Environment. Boston, MA: Harvard Business School Press (2000)
13. Stoeva, M.: Interactive Multimedia tool for dynamic generation of web interfaces with HTML5/PHP/MySQL and JavaScript. International Journal of Emerging Technology & Advanced Engineering, Vol. 4, Iss. 9, pp. 412–418 (2014)
14. Repository with contemporary tools for prototyping, project planning and Gantt charts, and sharing information, including, but not limited to: moqups.com, www.uxpin.com, docs.google.com, www.smartsheet.com (2017)
15. e-CF (European e-Competence Framework) version 3.0, adopted in 2016 as European standard and published by CEN as the European Norm EN 16234-1, http://www.ecompetences.eu/
16. Stoeva, M., Sharkov, G., Specialized page of "Software Quality Assurance (Q.A.)" course, http://edesign-bg.com/quality-software-2017.html (2017)

Author Index

Angelova Vesela 170

Avdjieva Irena 152

Bocevska Andrijana 28

Dimitrov Vladimir 15, 62, 70, 78, 97, 120

Genova Lyubka 158

Hadzhimateva Dilyana 152

Kaloyanova Elitsa 40

Kaloyanova Kalinka 48

Koleva Maria 125, 134

Kotevski Zoran 28

Krastev Evgeniy 106

Manevska Violeta 58

Muaremi Valmir 28

Najdovski Blagojce 58

Nikolova Yana 78

Nisheva-Pavlova Maria 7

Patias Ioannis 22, 89

Petrov Teo 48

Piredu Luca 78

Ristevski Blagoj 28

Savoska Snezana 28, 58

Savov Svetoslav 97

Sharkov George 184

Simeonova Valeriya 152

Stanev Ivan 125, 134

Stoeva Maya 184

Trifonova Mariana 178

Vassilev Dimitar 78, 97, 152