Information Systems & Grid Technologies

Sixth International Conference ISGT'2012

Sofia, Bulgaria, June 1–3., 2012.

ST. KLIMENT OHRIDSKI UNIVERSITY PRESS

ISGT'2012 Conference Committees

Co-chairs

- Prof Ivan SOSKOV
- Prof Kiril BOYANOV

Program Committee

- Míchéal Mac an AIRCHINNIGH, Trinity College, University of Dublin
- Pavel AZALOV, Pennsylvania State University
- Marat BIKTIMIROV, Joint Supercomputer Center, Russian Academy of Sciences
- Marko BONAČ, Academic and Research Network of Slovenia
- Marco DE MARCO, Catholic University of Milan
- Milena DOBREVA, University of Strathclyde, Glasgow
- Viacheslav ILIN, Moscow State University
- Vladimir GETOV, University of Westminster
- Jan GRUNTORÁD, Czech Research and Academic Network
- Pavol HORVATH, Slovak University of Technology
- Seifedine KADRY, American University of the Middle East, Kuwait
- Arto KARILA, Helsinki University of Technology
- Dieter KRANZMUELLER, University of Vienna
- Shy KUTTEN, Israel Institute of Technology, Haifa
- Vasilis MAGLARIS, National Technical University of Athens
- Violeta MANEVSKA, University of Bitola „St. Kliment Ohridski"
- Ivan PLANDER, Slovak Academy of Science
- Dov TE'ENI, Tel-Aviv University
- Stanislaw WRYCZA, University of Gdansk
- Fani ZLATAROVA, Elizabethtown College

Organizing Committee

- Vladimir DIMITROV
- Maria NISHEVA
- Kalinka KALOYANOVA
- Vasil GEORGIEV

Vladimir Dimitrov (Editor)

Information Systems & Grid Technologies

Sixth International Conference ISGT'2012

Sofia, Bulgaria, June 1–3., 2012.

Proceedings

St. Kliment Ohridski University Press

Preface

This conference was being held for the sixth time in the beginning of June, 2012 in the mountain resort Gyolechica near Sofia, Bulgaria. It is supported by the National Science Fund, by the University of Sofia "St. Kliment Ohridski" and by the Bulgarian Chapter of the Association for Information Systems (BulAIS). The Organizing Committee consists of scientists from the Faculty of Mathematics and Informatics of the University of Sofia. Traditionally this conference is organized in cooperation with the Institute of Information and Communication Technologies of the Bulgarian Academy of Sciences.

Total number of papers submitted for participation in ISGT'2012 was 53. They undergo the due selection by at least two of the members of the Program Committee. This book comprises 36 papers of 35 Bulgarian authors and 15 foreign authors included in one of the three conference tracks. In order to facilitate access to the results reported in this conference the papers with higher referee's score will be uploaded at AIS (Association of Information Systems) chapter "Conferences". 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 Editor

© 2012 Vladimir Dimitrov (Eds.)
ISSN 1314-4855
St. Kliment Ohridski University Press

TABLE OF CONTENTS

INFORMATION SYSTEMS

1. Tendencies in Data Warehouse Scope Development
Zlatinka Kovacheva 11

2. Relational Data Model as Evolution
Ivan Stanev 22

3. Dependencies and Interrelations of Courses in an Undergraduate Curriculum in Information Systems
Pavel I. Pavlov 28

4. Web-based System for Teamwork Evaluation
Radoslava Hristova, Goran Goranov, Elena Hinova 38

5. System Analysis of Information Systems for Local Economic Development Modeling – Case Study for the Region Pelagonia in R. of Macedonia
Snezana Savoska, Branko Dimeski 44

6. A Comparison of Change Management Implementation in ITIL, CMMI and Project Management
Kalinka Kaloyanova, Emanuela Mitreva 55

7. Modeling Z-specification in UML
Snezana Savoska 65

8. Logical Design for Configuration Management Based on ITIL
Hristo Kyurkchiev, Kalinka Kaloyanova 79

9. An Overview of the Moodle Platform
Vladimir Dimitrov 90

10. Planned Improvements for `moodle.openfmi.net`
Atanas Semerdzhiev 94

11. Choosing Approach for Data Integration
Hristo Hristov 98

12. Mining Bibliographic Data
Tsvetanka Georgieva-Trifonova 114

13. Build a model to predict the success of students by hereditary and social factors using the tool for data mining Weka
Jasmina Nedelkoska 125

14. Initial Results of a Plagiarism Detection System
Atanas Semerdzhiev 134

15. An Overview of the Department of Defense Architecture Framework (DoDAF)
Vladimir Dimitrov 139

INTELLIGENT SYSTEMS

16. Digital Spaces in "Popular" Culture Ontologized
Mícheál Mac an Airchinnigh 145

17. Application of knowledge management information systems in digital redactions
Elena Miceska 154

18. The Naïve Ontologist and "Intelligent" Story Telling
Mícheál Mac an Airchinnigh 163

19. AnatOM – An intelligent software program for semi-automatic mapping and merging of anatomy ontologies
Peter Petrov, Nikolay Natchev, Dimitar Vassilev, Milko Krachounov, Maria Nisheva, Ognyan Kulev 173

20. Usage of E-learning for Improving of Knowledge Management in Organizations
Vesna Mufa, Violeta Manevska, Natasha Blazeska-Tabakoska 188

21. ASP.NET-based Tool for Searching in Folklore Lyrics Stored in XML Format
Dicho Shukerov 196

22. Influence of Different Knowledge Management Techniques and Technology on Organization Effectiveness
Natasha Blazeska-Tabakovska, Violeta Manevska 206

23. IT Aspects of the Application of Ontologies in Academic Digital Libraries
Daniela Kjurchievska 213

24. A Method for Decision Making in Computer Science Research
Neli Maneva 220

DISTRIBUTED SYSTEMS

25. Digital Libraries and Cloud Computing
Maria M. Nisheva-Pavlova 231

26. Performance Evaluation of the Schemes for Dynamic Branch Prediction with ABPS Simulator
Milco Prisaganec, Pece Mitrevski, Nikola Rendevski 240

27. Data Analysis for Next Generation Sequencing – Parallel Computing Approaches in *de Novo* Assembly Algorithms
Dimitar Vassilev, Valeria Simeonova, Milko Krachunov, Elena Todorovska, Maria Nisheva, Ognyan Kulev,
Deyan Peychev, Peter Petrov, Dimitar Shiyachki, Irena Avdjieva 249

28. Requirements for Cloud Service Discovery Systems
Georgi Pashov, Kalinka Kaloyanova 280

29. Monitoring of Business Processes in the EGI
Radoslava Hristova 294

30. Computational Challenges in a Metagenomic Data Processing Pipeline
Milko Krachunov, Dimitar Vassilev, Ivan Popov, Peter Petrov 302

31. Simulation of the Behavior of a Pure Imperative Synchronous Programming Language by Means of Generalized Nets with Stop-Conditions
Magdalina V. Todorova 312

32. Declarative Semantics of the Program Loops
Krassimir Manev,

Trifon Trifonov 326

33. Blueprint of an Experimental Cloud Framework
Radko Zhelev 338

34. Adaptive Integrated Business Management Systems
Milko Tipografov, Evgeniya Grigorova 346

35. Algorithm for Simulating Timestamp Ordering in Distributed Databases
Svetlana Vasileva 352

36. Anycast DNS System in AS5421
Hristo Dragolov, Vesselin Kolev, Stefan Dimitrov 365

AUTHOR INDEX 375

Information Systems

Tendencies in Data Warehouse Scope Development

Zlatinka Kovacheva,

Higher College of Telecommunications and Post, Sofia, Bulgaria
Higher College of Technology, Muscat, Oman,
zkovacheva@hotmail.com

Abstract. A data warehouse is a repository of an organization's electronically stored data. Data warehouses are designed to facilitate reporting and analysis. The Data Warehouse environment positions a business to utilize an enterprise-wide data store to link information from diverse sources and make the information accessible for a variety of user purposes, most notably, strategic analysis. Data Warehousing requires both business and technical expertise and involves many activities. In the beginning it is necessary to identify the business information and to prioritize subject areas to be included in the Data Warehouse. This paper concerns the management of the scope of DataWarehouse.

Keywords: data warehouse, data mart, scope, gathering data, operational data, filtering, subject orientation

1 Introduction

Data warehousing began in the 1980s as a response to gather the information provided by the many application systems that were being built. Online applications served the needs of a limited community of users, and they were rarely integrated with each other. Additionally, online applications had no appreciable amount of historical data because they jettisoned their historical data as quickly as possible in the name of high performance. Thus, corporations had lots of data and very little information. Data warehousing began as a way to reduce users' frustration with their inability to get integrated, reliable, accessible data.

Usually Data warehouse is defined as a collection of data that is subject oriented, integrated and time variant for the purpose of management's decision processes [3,4]. Since the 1980s, data warehousing has gone from being a concept that was derided by the database theoreticians to conventional wisdom. Once the idea of the data warehouse became popular, vendors and consultants latched onto the concept as a good way to sell their products.. As a result, there was much confusion over what was and was not a data warehouse. Nowadays, people built different manifestations of data warehouses which fulfilled a real need in the marketplace.

The data in the Warehouse comes from the operational environment and external sources. Data Warehouses are physically separated from operational

systems, even though the operational systems feed the Warehouse with source data. All data in the data warehouse is identified with a particular time period.

A data warehouse was the first attempt at architecture that most organizations had ever encountered. Prior to data warehousing, everything had been a new application; however, it became apparent that applications were not going to get the organization where it needed to go over time. The solution was to build an architecture or at least the first fledgling steps of an architecture.

Data Warehouse applications are designed primarily to support executives, senior managers, and business analysts in making complex business decisions. Data Warehouse applications provide the business community with access to accurate, consolidated information from various internal and external sources [10].

The Data Warehouse process guides the development team through identifying the business requirements, developing the business plan and Warehouse solution to business requirements, and implementing the configuration, technical and application architecture for the overall Data Warehouse. It then specifies the iterative activities for the cyclical planning, design, construction, and deployment of each project [8].

One of the first stages of this process is the establishment of the scope of the Warehouse and its intended use. To paraphrase data warehousing author W. H. Inmon [6,7], traditional projects start with requirements and end with data. Data warehousing projects start with data and end with requirements. Once warehouse users see what they can do with the technology, they will want much more. That's why the Data Warehouse scope development is a never ending process.

2 Requirements Gathering

2.1 Source-Driven Requirements Gathering

Source-driven requirements gathering is a method based on defining the requirements by using the source data in production operational systems. This is done by analyzing an ER model of source data if one is available or the actual physical record layouts and selecting data elements deemed to be of interest. The major advantage of this approach is that it is known from the beginning what data can be supplied because it depends on the availability. Another benefit is that it minimizes the time required by the users in the early stages of the project. Of course, by minimizing user involvement, the risk of producing an incorrect set of requirements is increased. Depending on the volume of source data, and the availability of ER models for it, this can also be a very time-consuming approach. Perhaps, some of the user's key requirements may need data that is currently unavailable. Without the opportunity to identify such requirements, there is no

chance to investigate what is involved in obtaining external data from outside the organization. Even so, external data can often be of significant value to the business users [1].

A comparison between operational data and Data Warehouse data is given on Table 1[2,4].

Table 1. Operational vs. Data Warehouse data

Main characteristics	Operational data	Data Warehouse data
Type	Current, transient	Historical, periodic
Form	Raw, detailed, not normalized	Summarized, normalized
Quality	Inconsistencies and errors are included	Quality controlled – accurate with full integrity
Scope	Restricted	Comprehensive
Access	Specialized	Flexible
Update	Many updates	Periodic updates
Usage	Run the business on a current basis	Support managerial decision making

Although similar in nature to modeling and design in the operational world, the actual steps in data warehousing are different. Operational models are typically ER models and the data warehousing models are dimensional models. The data warehouse model looks more physical in nature than a model of an operational system. Probably the feature that most differentiates the data warehouse model from the logical operational model is the denormalization of the dimensions. They have to be organized for the purposes of management and suitable for the end users points of view to the business data.

The result of the source-driven approach is to provide the user with what is available. Relative to dimensional modeling, it can be used to drive out a fairly comprehensive list of the major dimensions of interest to the organization. This could minimize the proliferation of duplicate dimensions across separately developed data marts. Also, analyzing relationships in the source data can identify areas on which to focus your data warehouse development efforts.

2.2 User-Driven Requirements Gathering

User-driven requirements gathering is a method based on defining the requirements by investigating the functions the users perform. This is usually done through a series of meetings and/or interviews with users. The major advantage to this approach is that the focus is on providing what is needed, rather than what is available. In general, this approach has a smaller scope than the source-driven

approach. Therefore, it generally produces a useful data warehouse in a shorter time. The users must clearly understand that it is possible that some of the data they need can simply not be made available. If a user is too tightly focused, it is possible to miss useful data that is available in the production systems [1].

User-driven requirements gathering is the approach of choice, especially when developing data marts. For a full-scale data warehouse, both methods are applicable. It would be worthwhile to use the source-driven approach to break the project into manageable pieces, which may be defined as subject areas. The user-driven approach could then be used to gather the requirements for each subject area.

2.3 Requirements Gathering Problems

Usually around 80% of the time building a data warehouse has been spent on extracting, cleaning and loading data [12]. A lot of problems appear during the process of requirements gathering.

Some problems come from OLTP (On-line transaction processing) systems. Some of their errors can be fixed in the Data Warehouse. Sometimes the data in OLTP systems are not validated. Typically once data are in warehouse many inconsistencies are found with fields containing 'descriptive' information. For example, many times no controls are put on customer names. This is going to cause problems for a warehouse user who expects to perform an ad hoc query selecting on customer name. The warehouse developer, again, may have to modify the transaction processing systems or develop (or buy) some data scrubbing technology.

Usually OLTP systems are built by different software tools and the problem of comparability appears. In every big company there are many OLTP systems for the purposes of different activities, e.g. financial, billing, marketing, human recourses and other systems. They are developed by different departments of the company, using different software environments and very often they produce contradictions. The process of uploading the data from these systems into the data warehouse has to solve such conflicts.

A very common problem is to find the need to store data that are not kept in any transaction processing system. For example, when building sales reporting data warehouses, there is often a need to include information on off–invoice adjustments not recorded in an order entry system. In this case the data warehouse developer faces the possibility of modifying the transaction processing system or building a system dedicated to capturing the missing information [10].

Another problem is so called 'granularity' problem. Some transaction processing systems don't contain details. This problem is often encountered in customer or product oriented warehousing systems. Often it is found that

a system which contains information that the designer would like to feed into the warehousing system does not contain information down to the product or customer level.

In addition to understanding the feeder system data, it can be advantageous to build some of the "cleaning" logic on the feeder system platform if that platform is a mainframe. Often cleaning involves a great deal of sort/merging – tasks at which mainframe utilities often excel. Also, it is possible to build aggregates on the mainframe because aggregation also involves substantial sorting.

Very often, the end users increase their requirements when they started to use the data warehouse and realize its real advantages. It comes about because the query and report tools allow the user the users to gain a much better appreciation of what technology could do.

Many warehouse users define conflicting business rules or they don't know how to use the data even if they passed training courses. That's why periodically training of the users is required, to provide a contemporary view to the information of the updated data warehouse.

Large scale data warehousing can become an exercise in data homogenizing. A popular way to design a decision support relational databases is with star or snowflake schemas. Persons taking this approach usually also build aggregate fact tables. If there are many dimensions to the data, the combination of the aggregate tables and indexes to the fact tables and aggregate fact tables can use many times more space than the raw data. If multidimensional databases are used, certain products pre–calculate and store summarized data. As with star/snowflake schemas, storage of this calculated data can use far more storage than the raw data.

Security problems are also very important, especially if the data warehouse is Web-accessible. The information of the data warehouse can be uploaded and downloaded by many authorized users for different purposes. That's why the access to the data should be managed properly and the data on each level of summarization should be well protected.

Reorganizations, product introductions, new pricing schemes, new customers, changes in production systems, etc. are going to affect the warehouse. If the warehouse is going to stay up to date, changes to the warehouse have to be made fast. Customer management issues require a proper maintenance of the data.

3 Data Warehouse Scope Establishment as a Stage of the Data Warehouse Process

The Data Warehouse process is conducted in an iterative fashion after the initial business requirements and architectural foundations have been developed with the emphasis on populating the Data Warehouse with functional subject-area

information in each iteration. The process guides the development team through identifying the business requirements, developing the business plan and Warehouse solution to business requirements, and implementing the configuration, technical, and application architecture for the overall Data Warehouse. It then specifies the iterative activities for the cyclical planning, design, construction, and deployment of each population project.

The most important strategic initiative is analyzed to determine the specific business questions that need to be answered through a Warehouse implementation. Each business question is assessed to determine its overall importance to the organization, and a high-level analysis of the data needed to provide the answers is undertaken. A business question can be answered through objective analysis of the data that is available. The data is assessed for quality, availability, and cost associated with bringing it into the Data Warehouse. The business questions are then revisited and prioritized based upon their relative importance and the cost and feasibility of acquiring the associated data. The prioritized list of business questions is used to determine the scope of the first and subsequent iterations of the Data Warehouse, in the form of population projects. Iteration scoping is dependent on source data acquisition issues and is guided by determining how many business questions can be answered in a three to six month implementation time frame. A "business question" is a question deemed by the business to provide useful information in determining strategic direction.

3.1 Validating the Model

Before investing a lot of time and effort in designing the warehouse, it is absolutely necessary to validate the model with the end users. The purpose of such a review is twofold. First, it serves to confirm that the model can actually meet the users' requirements. Second, and just as important, a review should confirm that the user can understand the model. Once the warehouse is implemented, the user will be relying on the model on a regular basis to access data in the warehouse. No matter how well the model meets the users' requirements, the warehouse will fail if the user cannot understand the model and, consequently, cannot access the data. Validation at this point is done at a high level. This model is reviewed with the user to confirm that it is understandable. Together with the user, the model has to be tested by resolving how it will answer some of the questions identified in the requirements. It is almost certain that the model will not meet all of the users' requirements.

Either the requirements need to be better understood or, as is often the case, they have changed and need to be redefined. Usually, this will lead to additions, and possibly changes, to the model already created. In the mean time, the validated portion of the model will go through the design phase and begin providing benefits

to the user. The iteration of development and the continued creation of partially complete models are the key elements that provide the ability to rapidly develop data warehouses.

3.2 Identifying the Sources

Once the validated portion of the model passes on to the design stage, the first step is to identify the sources of the data that will be used to load the model. These sources should then be mapped to the target warehouse data model. Mapping should be done for each dimension, dimension attribute, fact, and measure. For dimensions and facts, only the source entities (for example, relational tables, flat files, IMS DBDs and segments) need be documented. For dimension attributes and measures, along with the source entities, the specific source attributes (such as columns and fields) must be documented.

Conversion and derivation algorithms must also be included in the metadata. At the dimension attribute and measure level, this includes data type conversion, algorithms for merging and splitting source attributes, calculations that must be performed, domain conversions, and source selection logic [9]. A domain conversion is the changing of the domain in the source system to a new set of values in the target. Such a conversion should be documented in the metadata.

In some cases target attributes are loaded from different source attributes based on certain conditions. At the fact and dimension level, conversion and derivation metadata includes the logic for merging and splitting rows of data in the source, the rules for joining multiple sources, and the logic followed to determine which of multiple sources will be used.

Identifying sources can also cause changes to the model. This will occur when a valid source is not available. Some time there is no source that comes close to meeting the users' requirements. This should be a very rare case, but it is possible. If only a portion of the model is affected, it is recommended to remove that component and continue designing the remainder. A better scenario happens when there will be a source that comes close but doesn't exactly meet the user's requirements. In this case it will be necessary only to modify slightly the model.

4 Some Tendencies in Data Warehouse scope Development

Data Warehouse is developed in an environment where technology could be changed as fast as the business requirements change. The Data Warehouse environment has features like life cycle, integration of structured and unstructured data, metadata, etc. In order to adapt to the ongoing change of the business, semantically temporal data (data that undergoes changes) should be separated from semantically static data, which does not change over time. When this

happens a new snapshot of the data is created. Each snapshot is delimited by time and a historical record for the semantically temporal data is created. In order to meet the changes in the environment, a new generation of Data Warehouse has been developed – DW 2.0 [5].

The analysis of the characteristics of DW 2.0 presents the basic differences between the two generations data warehouses. They cover crucial aspects of the data warehousing and span from the building methodology, through the development, to the implementation of the final product.

Several fundamental aspects underlie the difference between DW 2.0 and the first-generation data warehousing. The new features of DW 2.0 concerning the scope of the Data Warehouse are the following [6,7]:

4.1 The Lifecycle of Data

As data ages, its characteristics change. As a consequence, the data in DW 2.0 is divided into different sectors based on the age of the data. In the first generation of data warehouses, there was no such distinction.

According to the concept of "life cycle of data" data passes through four different sectors:

- The Interactive Sector
- The Integrated Sector
- The Near Line Sector
- The Archival Sector

The Interactive Sector is the place where the data enters the DW 2.0 environment either from an application outside DW 2.0 or as part of a transaction from an application which is located in the Interactive Sector. On leaving the Interactive Sector the data goes to the Integrated Sector. At this moment "data quality processing" like "domain checking and range checking" is done. Data is collected and kept with maximum time span of three to five years. After leaving the Integrated Sector the data can go either to the Near Line Sector or to the Archival Sector.

4.2 Structured and Unstructured Data

Another aspect of the DW 2.0 environment is that it contains not only structured but also unstructured data. Unstructured data is a valid part of the DW 2.0 [14]. Some of the most valuable information in the corporation resides in unstructured data. The first generation of data warehouses did not recognize that there was valuable data in the unstructured environment and that the data belonged in the data warehouse.

Two basic forms of unstructured data – "textual" and "non-textual" are

supported. Textual unstructured data could be found in emails, telephone conversations, spreadsheets, documents and so forth. Non-textual unstructured data occurs as graphics and images. The integration of structured and unstructured data in DW 2.0 enables different kinds of analysis - against unstructured data or a combination of structured and unstructured data.

Unstructured data exists in several forms in DW 2.0 - actual snippets of text, edited words and phrases, and matching text. The most interesting of these forms of unstructured data in the DW 2.0 environment is the matching text. In the structured environment, matches are made positively and surely. Not so with unstructured data. In DW 2.0, when matches are made, either between unstructured data and unstructured data or between unstructured data and structured data, the match is said to be probabilistic. The match may or may not be valid, and a probability of an actual match can be calculated or estimated. The concept of a probabilistic match is hard to fathom for the person that has only dealt with structured systems, but it represents the proper way to link structured and unstructured data.

4.3 A New Concept of Metadata

Metadata is the glue that holds the data together over its different states [11]. The first generation of data warehousing omitted metadata as part of the infrastructure.

Metadata is found in many places today - multidimensional technology, data warehouses, database management system catalogs, spreadsheets, documents, etc. There is little or no coordination of metadata from one architectural construct to another; however, there is still a need for a global repository. These sets of needs are recognized and addressed architecturally in DW 2.0.

In DW 2.0 every sector has metadata of its own. The metadata for the Archival Sector is placed directly with the archival data while the metadata for the other sectors is kept separately from the data. The different sectors support metadata for the structured and for the unstructured data.

4.4 Integrity of the Data

The need for integrity of data is recognized as data passes from online processing to integrated processing [13].

The DW 2.0 environment has features like integration of structured and unstructured data and metadata. In order the technology to adapt to the ongoing change of the business, semantically temporal data (data that undergoes changes) should be separated from semantically static data, which does not change over time. When this happens a new snapshot of the data is created. Each snapshot is delimited by time and has to date and from date. In this way a historical record

for the semantically temporal data is created.

4.5 The Spiral Methodology

One of the most prominent characteristics of DW 2.0 is the spiral methodology, which adopts the so called "seven streams approach". Different activities are initiated and completed in each stream [5]:

- The Enterprise reference model stream concerns the creation and continued maintenance of a corporate data model which consists of Entity Relationship Diagrams and background information;
- The Enterprise knowledge coordination stream relies on various artifacts that come out of the corporate data modeling, information factory development, and data profiling streams;
- The Information factory development stream is built "topic by topic";
- The Data profiling and mapping stream includes the analysis of the data from the source systems
  in terms of quality and completeness;
- The Data correction stream - the data is examined and decided which part of it will be corrected;
- The Infrastructure management stream addresses the decision what people resources and tools will be used;
- The Total information quality management stream provides data quality monitoring and process improvement.

5 Conclusion

The analysis of the characteristics of DW 2.0 presents the basic differences between the two generations data warehouses. They cover crucial aspects of the data warehousing from the building methodology, through the development, to the implementation of the final product.

The huge amount of unstructured and unprocessed data from different sources from one side and a large number of on-line transaction processing systems organized on different platforms from the other side have to be integrated uniformly in order to satisfy the rapidly increasing business requirements.

The tendencies concerning Data Warehouse Scope development are discussed in this paper. They follow the process of enlarging and complexness of the data stored nowadays, the increasing number of data sources of different type and the necessity of integration of the data in order to provide more powerful information for the purposes of the management, marketing, financial and all other activities in the fast growing and rapidly changing business environment.

References

1. Ballard Ch., Herreman D., Schau D., Bell R., Kim E., Valencic A.: Data Modeling Techniques for Data Warehousing International Technical Support Organization, February 1998, http://www.redbooks.ibm.com ,SG24-2238-00 IBM
2. Browning D., Mundy J.: Data Warehouse Design Considerations, Microsoft® SQL Server™ 2000, December 2001
3. Chaudhuri, Surajit and Dayal, Umeshwar: An Overview of Data Warehousing and OLAP Technology, ACM New York, NY, USA (1997)
4. Hoffer J. A. , Prescott M. B., McFadden F.R.:2 Modern Database Management, 7th Edition, Prentice Hall, (2005)
5. Hristov N., Kaloyanova K.: Applying a Seven Streams Approach for Building a DW 2.0 , Proceedings of the Fourth International Conference On Information Systems & Datagrids, Sofia, Bulgaria, May 2010
6. Inmon, William and Strauss, Derek and Neushloss, Genia: DW 2.0: The Architecture for the Next Generation of Data Warehousing, Elsevier Inc. (2008)
7. Inmon, Bill: Different Kinds of Data Warehouses. http://www.b-eye network.com/view/10416, (2009)
8. Naydenova I. , Kaloyanova K.: Basic Approaches Of Integration Between Data Warehouse And Data Mining,
   Proceedings of the First International Conference On Information Systems & Datagrids , Sofia, Bulgaria, (2005), http://isgt.fmi.uni-sofia.bg/index.php?contentPage=proceeding
9. Naydenova I. , Kaloyanova K.: An Approach Of Non-Additive Measures Compression In Molap Enviroment", Proceedings of the IADIS Multi Conference on Computer Science and Information Systems, 2007, 3 - 8 July 2007, Lisbon, Portugal, ISBN: 978–972–8924–35-5,pp. 394-399, http://www.mccsis.org/2007/
10. Naydenova I. , Kaloyanova K., Ivanov St.:A Multi-Source Customer Identification", Proceedings of the Third International Conference On Information Systems & Datagrids, Sofia, Bulgaria, pp.77-85, 28 - 29 May 2009
    http://ci.fmi.uni-sofia.bg/isgt2009/
11. Poole J.: Model-Driven Data Warehousing, Hyperion Solutions Corporation, Burlingame, CA, January 2003
12. Rahm, Erhard and Do,H. H.: Data Cleaning: Problems and Current Approaches, (2000)
13. Russom, Ph.: Complex Data: A New Challenge for Data Integration, (2007)
14. Santos, Ricardo Jorge, Bernardino, Jorge: Optimizing Data Warehouse Loading Procedures for Enabling Useful-Time Data Warehousing, (2009)

Relational Data Model as Evolution

Ivan Stanev,

University of Ruse, 8 Studentska Str., 7017 Ruse, Bulgaria
ins@ait.ru.acad.bg

Abstract. Relational Data Model is frequently accepted as revolution in data modeling. The aim of this paper is to show that it is not true. Relational Data Model is a result of natural evolution based on previous experience and investigations on data modeling.

Keywords: data base, relational data model, data modeling.

1 File Organization

For the first time, relational model of data was published in [1]. Codd intention was to solve some problems with currently used data models. It is better to take an excursion in the past preceding relational model.

In the last 70-ies, first Data Base Management Systems (DBMS) had emerged and many new problems arose in the area of data modeling. DBMS are result of many years, research and investigation on data representation and storage in computing systems.

The most important invention of these days was magnetic disks as primary media for secondary storage. This invention traced then the development of 'file' concept. Initially, the programmer was responsible data placement on the disk sectors and their retrieval. Latter on this responsibility was attached to the operating system. At first, this functionality was implemented as utility programs for file cataloging and disk space management, but then it became part of the operating system. Such an example is the leading these days operating system IBM DOS/360 (Disk Operating System for System 360). For more details on IBM DOS/360 refer [2]. Now the programmer is able to work with files directly supported by the operating system that got responsibility to manage the files: catalog files, pack/unpack records in/from blocks, and allocate/release blocks. Files of records were a new level of abstraction based on the underlying hardware system. Here is the place to mention, that files of bytes (as used in programming languages like C) are abstraction too on the underlying hardware system, but this concept is used to support different kinds of information media. A file of bytes could be placed on magnetic disk, magnetic tape, punched tape, stack of punched cards and so other media. File of bytes is naturally sequentially accessed

V. Dimitrov (Editor): ISGT'2012. ISSN 1314-4855
Proceedings of the 6th International Conference on
Information Systems and Grid Technologies, Sofia, June 1-3., 2012.

because all kind of media support sequential read/write. Files of records are naturally randomly accessed. This does not means that file of records could not be sequentially accessed. Files of records with random access are an abstraction of files of sectors on disks. Direct access to records could be supported on magnetic tapes, but it is not efficient.

In random access files, records are ordered in sequence and every record is accessed by its number. This is an abstraction of cataloged file as chain of blocks. The operating system manages a catalog with information about every file and its chain of blocks. It is responsible for consistency of data in the file, i.e. a block is not allocated to two files and data could not be corrupted by elementary errors as it has been earlier. The operating system transparently allocates and releases blocks to/from files. The blocks in these files are accessed directly by their numbers. A block is a sequence of several sectors. Sector is the elementary read/write unit in magnetic disks. Block is used as a concept in the operating systems for optimization purposes.

Records in files of records are accessed directly by their numbers. This is as in files of blocks and it is the most important concept of files of records. There are some problems related with the files of varying length records. These problems are related to record modification. When a record is updated and its new length is greater than the old length, where the new record could be placed? This new record could not be placed on the old one.

The above mentioned problem could be solved with a new abstraction. Instead identifying records by their sequence numbers a key could be used for this purpose. In this case, random access to the files is based on the keys that uniquely identify records in the file. In such a way records are unpinned from their positioning in the file. With concept of key many file organizations have been developed like indexed-sequential files, B-trees, hash files etc. Even some of them became a language construct in some of the early high level programming languages lake PL/1 and Cobol [3,4].

2 Database Management Systems

The next level of abstraction is based on the file organization. When data is stored in some file organization, this means that some concrete record ordering based on a key is applied. When the file organization for the date is applied, then another ordering by key is applied. This means that an application program that uses some key ordering would fail if it is changed. This is a problem.

First DBMS were based on hierarchy and network organizations of data such systems are IBM IMS [5] and IDMS [6]. They continued to support the concept of record ordering. They offered many utilities for efficient access to records, but their concepts remained in the area of engineering solutions based

on file organization. In these DBMS, navigation is based on record ordering and pointers. These DBMS are evolutionary step because their data models do not prescribe specific implementation how this is done in file organizations.

At first glance, relational model do not participate in this evolution. It is based on mathematical theory of sets; it is not further abstraction of some elements of file organization. But relational model is a jump in the evolution of data representation. It is further abstraction of data organization and navigation.

Codd had been motivated to develop a new model of data independent of application programs and way of data access. Now, data access efficiency became DBMS responsibility. Application program, on the other hand, has more simplified and more abstract interface to the date through DBMS. Relational model represents data in their natural structuration without any elements of their machine implementation, how this has been done in previous models of data. Navigation in relational model is available through high level query language. In such a way, application program only declares needed navigation functionality and DBMS implements it.

Relational model is evolution (more precisely jump in the evolution), because Codd intention was to develop new model of data to overcome pitfalls of currently available models of data. The main target was to represent the data to applications in an independent way of its storage. There three dependencies that relational model had to fight: ordering, indexing and access path dependencies. In the next, a brief discussion on their nature is given.

Ordering dependence. Elements of data could be stored in the data in different ways: in some cases, ordering is not important; in other cases, the element could participate in only one ordering; or the element could participate in several orderings. If an application uses some ordering of data and this ordering is changed, then the application will fail. The pre-relational models do not differentiate representation ordering from storage ordering of data.

Indexing dependence. Creation of indexes on the data permits to accelerate access to data, but it decreases data modifications: update, insert and delete operations. In environments with changeable characteristics of work with data, it is hazardous to use indexes. The problem is the impact of dynamic index creation/removal on applications whose access to the data uses indexes. DBMS administration of index management can solve this problem – not application programmer.

Dependence of access paths. In pre-relational models of data, navigation must follow the hierarchy or graph structuring of data. This means that application programs are bound to the concrete structuring of data. So, every change in data structure could result in some application program fail.

The intention of relational model is to solve above mentioned problems – it is not abstract mathematical construction.

3 Relational Model of Data

Relational model is based on the concept 'relation' from mathematical theory of sets. Relation is a subset of Cartesian product of several domains. Relation is represented as an array with columns named by the domains participating in the Cartesian product. This array has as many rows as tuples has the relation. Every column of this array contains values only from the domain that names it. Every row corresponds to a tuple of the represented relation.

Codd names columns by the domains. This is problem when a domain is used more than one time in the Cartesian product. To resolve this ambiguity, Codd adds a role to distinguish usage of such domains. Using domains (and roles), Codd tries to give way for specification of data semantic.

Formally, relational model is presented in [7], where this initial view on the relations is discussed in details.

Initial implementations of relational model showed deviations from presentation in [1]. For example, semantics of domains is the corner stone of semantics of relations, but these first DBMS do not have mechanisms for complex domain management, instead they have as domains the sets of types of conventional programming languages. The last ones are taken from mathematics and they are very abstract to contain any suitable semantics for data modeling of the real world. That is why Codd, later on, tried several times to redefine relational model to capture more meaning as in [8].

In relational model, data structures are independent of ordering. Relation is a set of tuples and navigation is independent of their ordering. There is a freedom to represent the same relation with different file organizations.

One of the questions that arise is why Codd has used concept of relation as fundamental structure in the model instead of more generalized concept of set? In reality, the relation is a specialized kind of set. In many models of data basic element is the set. Codd's selection has deep roots in theory and practice. First, relation is an abstraction of data storage on disks. There, data are stored in atomic units – sectors with fixed length. The sector is the smallest unit for read/write operations on the disk. For optimization purpose could be used block that are multiple sequential sectors, but this does not change the situation. Relation could be viewed as a machine representation of data on the disk where relation corresponds to disk and tuples – to sectors.

Every domain has some representation of its values. This representation could be fixed or varying one. Tuples could be of fixed or varying length depending on representation of relation domains values. Varying length tuples are represented with fixed length blocks. Efficiency of access to tuples depends of efficiency of access to fixed length blocks. Tuples are abstraction of the records from underlying file system. Tuples do not put any restrictions on the record representation in

the different file organizations. This means that relation representation could be freely changed from one file organization to another.

Relational model is in reality abstraction of magnetic disk structure. From the time of relational model emergence till now, there is no basic change in the concept of secondary storage media – magnetic disks. That is why relational model is still alive.

There are some alternative implementations of relational model that are not based the abstraction tuple-record and relation-file. For example, some of these implementations try to represent relation as relationship among its domains; the values are stored in domain structure; and the relation structure does not contain domain values, but only pointers to them. This approach does not follow the natural evolution of data representation and has not received industry support.

A variation on relational model is binary model of data [10], where all relations are binary ones. This approach again is deviation from the evolution and has not received industry support.

So, the relation, how was introduced from the beginning, was an abstraction of file organization with random access by key. This relation has one dimension of freedom – independence of tuple ordering, but still there is a problem with the fixed domain ordering. This is a problem from user's point of view – it is not implementation problem. The user has to know relation name, the names of its domains (and roles for duplicated domains) and domains ordering. Codd solved this problem introducing 'relationships'. This is a relation in which domain ordering is not important. Relationship represents all relations that could be produced by all permutations of its domains. Now, the user has to know relationship name and its domain names. But even this information could be derived with some DBMS utilities. Relationship is closer to nowadays term 'table'. More detailed discussion and formal specification can be found in [11]. In this specification the database is a set of relationships implemented with a database of relations how the things happen to be in the implementations. So, from users point of view DBMS manages relationships, but from implementation point of view – relations.

In this presentation of relational model, there are no indexes, pointers or access paths. Instead, there is a high level query language – relational algebra. DBMS must implement this query language to support relational model. Relational algebra is an abstraction of navigation without any access paths, indexes and pointers. It is based on set operations with some extensions. This is a declarative query language – not navigational one. The application program has to declare its data needs without specifying any access paths or indexes. DBMS job is to retrieve and return these data. Access paths and indexes are part of DBMS implementation – not of relational model. The application program is not bounded with any details on the navigation. Relational model does not depend on access paths and indexes. Relational algebra on relations is formally specified in

[12]. The operations in this specification are derived from [1].

4 Conclusion

Relational model is a result of natural evolution in data modeling. It is a jump in the evolution that has deep roots in previous data models and underlying hardware structure. That is why relational model is still on the stage.

References

1. Codd, E.F.: A Relational Model of Data for Large Shared Data Banks. CACM vol. 18, no 6, 377--387 (1970)
2. DOS/360 and successors: http://en.wikipedia.org/wiki/DOS/360_and_successors
3. PL/I: http://en.wikipedia.org/wiki/PL/I
4. COBOL:, http://en.wikipedia.org/wiki/COBOL
5. IBM Information Management System: http://en.wikipedia.org/wiki/IBM_Information_Management_System
6. IDMS: http://en.wikipedia.org/wiki/IDMS
7. Dimitrov, V.: Formal Specification of Relational Model of Data in Z-Notation, Proc. of XXXIX Conf. of UMB, 178-183 (2010)
8. Codd, E.F.: Extending the Database Relational Model to Capture More Meaning, ACM Transactions on Database Systems, Vol. 4, No. 4, 397—434 (1979)
9. Seshadri, P., Livny, M., Ramakrishnan, R.: SEQ: Design and Implementation of Sequence Database System, Proc. of the 22nd Int. Conf. on Very Large Data Bases (VLDB '96), Mumbay, India, 99—110(1996)
10. Abrial, J.R.: Data Semantics, Proc. of IFIP Conf. on Data Base Management, North-Holland, 1—59 (1974)
11. Dimitrov, V.: "Relationship" Specified in Z-Notation, Physics of Elementary Particles and Atomic Nuclei, Letters, Vol. 8, No. 4(167), 655—663 (2011)
12. Dimitrov, V.: Formal Specification of Relational Model with Relational Algebra Operations, Proc. of Fifth Int. Conf. ISGT, 25-28 May 2011, Sofia, pp. 57-74.

Dependencies and Interrelations of Courses in an Undergraduate Curriculum in Information Systems

Pavel I. Pavlov

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

Abstract. Information Systems is an academic field which ranges over two broad areas: (1) functionalities of information systems (acquisition, deployment and management of information technology resources and services) and (2) development and evolution of technology infrastructures. The architecture of the IS 2002 model curriculum for undergraduate degree programs in Information Systems, developed due to the joint efforts of AIS, ACM and AITP, consists of five curriculum presentation areas: information systems fundamentals; information systems theory and practice; information technology; information systems development; and information systems deployment and management processes. The paper presents an analysis of the body of knowledge covered by three specific courses within the undergraduate program in Information Systems at the Faculty of Mathematics and Informatics, Sofia University. These courses are concrete implementations of some learning units included in IS 2002.9 "Physical Design and Implementation with Programming Environments". The main goals and objectives of these courses, their interrelations and underlying technologies are discussed as well.

Keywords: IS 2002 model curriculum, data definition language, schema definition language, active server pages, e-business systems, XML, ASP.NET.

1 Introduction

Extensible Markup Language (XML) is a markup language that defines a syntax for encoding documents in a format that is both human-readable and machine-readable. The design goals of XML include generality, simplicity and interoperability. Although the purpose of XML is focused on documents, it is widely used for the representation of arbitrary data structures, for example in Web technology standards and Web services.

Active Server Pages (ASP) was Microsoft's first server-side script engine for dynamically generated Web pages. Subsequently it was superseded by ASP.NET. The latter may be characterized as a Web application framework developed to allow programmers to build dynamic Web sites, Web applications and Web services. In particular, the combination of ASP.NET and XML-based Web services presents

V. Dimitrov (Editor): ISGT'2012. ISSN 1314-4855
Proceedings of the 6th International Conference on
Information Systems and Grid Technologies, Sofia, June 1-3., 2012.

an appropriate technological base for building E-Business applications. The paper discusses the educational objectives and the contents of three interrelated courses taught at the Faculty of Mathematics and Informatics, Sofia University: XML Programming, ASP Programming and E-Business Systems.

2 XML Programming Course

The course is aimed at introduction to XML – a markup language, used as a standard for description of the structure of data in meaningful ways. Anywhere that data is input/output, stored, or transmitted from one place to another, is a potential fit for XML's capabilities.

XML (eXtensible Markup Language) [1, 2] is an open standard developed by W3C (World Wide Web Consortium). It supports two interrelated types of applications: 1) document encoding aimed at publication on the Internet and 2) exchange of data. A significant feature of XML is that it dissociates the encoding of data from the program logic and the user interface's code. This is a stepping-stone to platform independence and reusability of resources.

XML is based on the Standard Generalized Markup Language (SGML) [3] which is the basis of various works in the area of electronic transfer of documents by definition of sets of tags, setting up the corresponding document type definition (DTD).

Course projects in XML Programming

The course projects in XML Programming are tasks in which students have to construct a model of document. The document describes properties and attributes of a real phenomenon or object – proper examples are football matches, cars, films, computers, weapons, specific sports, etc.

Each student chooses and develops his/her own model of XML document describing a chosen application area. A Document Type Definition (DTD) or Schema is defined for this model. The DTD (Schema) should contain minimum 20 different elements and 5 attributes which determine the contents and the structure of the model.

After that, one creates 5 different instances (XML files) of this document using a text editor or a special-purpose XML editor. The text content of the documents is in Bulgarian language, written in Cyrillic.

A well-formed and valid XML file should be created by DOM (by using proper script language – JavaScript or Visual Basic Script). The texts of elements and values of attributes are contained in the corresponding programs. The resulting HTML page is expected to be demonstrated in a browser.

Moreover, 4 different XSLT files for transformation of XML documents to HTML files, 2 different XSLT files for transformation of XML documents to

other XML documents, and 2 different XSLT files for transformation of XML documents to plain text, have to be created. The newly generated HTML files, XML documents and plain text are demonstrated in a browser.

3 ASP Programming Course

The goal of the course is to introduce to the main technics at building of active pages. ASP (Active Server Pages) is an Internet technology developed by Microsoft. The basis of ASP consists of the interpretation of program code sent by a client to a server. In static HTML pages additional elements interpreted by the server are added. The server processes the script code from the client page and the code is converted to HTML format before returning to the client's browser. It is expected from the participants in course to have knowledge about WWW, HTML, SQL and skills for building Web pages.

ASP.NET [4, 5] came later as a subset of .NET Framework. ASP.NET was developed as an industrialstrength Web application framework that could address the limitations of ASP. Compared to classic ASP, ASP.NET offers better performance, better design tools, and a rich set of readymade features.

ASP.NET is an object-oriented programming model which enables one to compose a Web page as easy as to create a Windows application.

ASP.NET is designed as a server-side technology. All ASP.NET code executes on the server. When the code is finished executing, the user receives an ordinary HTML page, which can be viewed in any browser. Figure 1 illustrates the difference between the server-side model and the client-side one.

These are some other reasons for avoiding client-side programming:

- *Isolation*. Client-side code can't access server-side resources. For example, a client-side application has no easy way to read a file or interact with a database on the server (at least not without running into problems with security and browser compatibility).
- *Security.* End users can view client-side code. And once malicious users understand how an application works, they can easy tamper with it.
- *Thin clients.* As the Internet continues to evolve, Web-enabled devices such as mobile phones, palmtop computers, and PDAs (personal digital assistants) are appearing. These devices can communicate with Web servers, but they don't support all the features of a traditional browser. Thin clients can use server-based Web applications, but they won't support client-side features such as JavaScript.

However, client-side programming isn't truly inapplicable. In many cases, ASP.NET allows one to combine the best of client-side programming with server-side programming. For example, the best ASP.NET controls can intelligently detect the features of the client browser. If the browser supports JavaScript, these

controls will return a Web page that incorporates JavaScript for a richer, more responsive user interface.

Course projects in ASP Programming

The course projects in ASP programming are tasks which require the use ASP. NET technology and C#. First of all, the student has to construct a model of document. The document will describe some properties and attributes of a real phenomenon or object – e.g. football games, cars, films, computers, weapons, specific sports, etc. The text content of the document should be in Bulgarian language, written in Cyrillic.

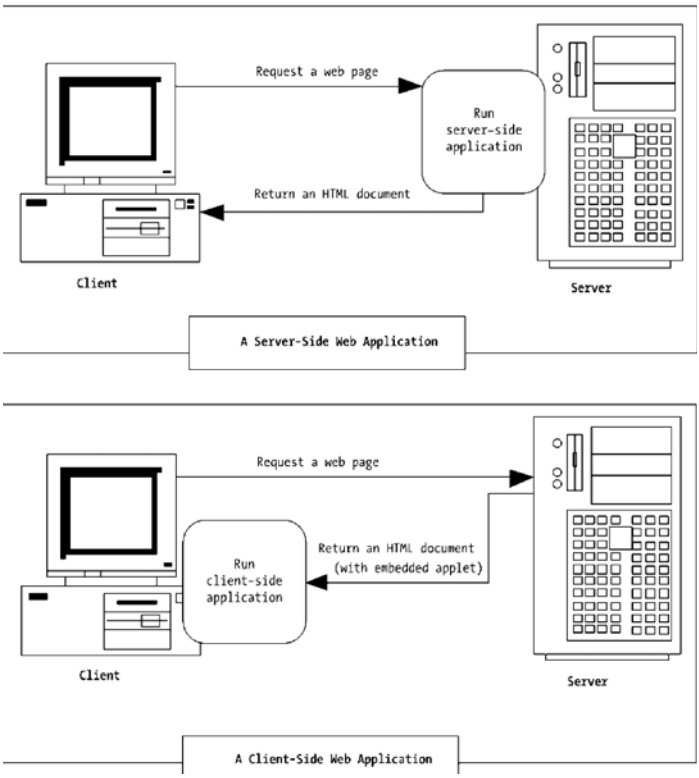


Fig. 1. Server-side and client-side Web applications

Each student chooses and develops his/her own model of XML document describing a chosen area of interest. A DTD or Schema has to be defined for this model. The DTD (Schema) should contain minimum 30 different elements and 10 attributes which determine the contents and the structure of the model.

Then one creates 20 different instances (XML files) of the model document using a text editor or a special-purpose XML editor.

As a next step, a database (DB) in Microsoft SQL Server 2008 is created, that corresponds by content to the content of the XML documents.

From a proper menu a mode of action is chosen, at which the content of all 20 XML files is transferred to the DB. Before transference the content of each file is checked for validity against the DTD (Schema). The result of checking (yes/no) is printed along with the file name.

Then, from a proper menu, a mode of action is chosen, at which the content of each particular XML document is assigned to the fields of a corresponding form in the sense of HTML. After filling the fields of the form, their content is transferred to the DB after validity checking. Besides of this a XML file is created and recorded.

4 The Course in E-Business Systems

The course in E-Business Systems is intended for software developers responsible for designing and building Internet-based applications for the .NET platform. It provides detailed information on implementing a .NET system design, and presents the best practices and lessons learned along the way. The course is also appropriate for information system managers faced with planning issues. It explains the key concepts behind the most important technologies in the .NET platform. Visual C# .NET, Visual Basic .NET, ASP.NET, ADO.NET, SOAP, and XML-based Web Services are all covered having in mind the specific goals of the course. Some practical examples of how to effectively implement these technologies have been provided. The design and implementation of an Internet-based E-Business application has been scheduled as well.

This course is planned as a continuation of the ASP Programming and XML Programming courses for students at the Faculty of Mathematics and Informatics. The participants are expected to have knowledge and skills in WWW, HTML, SQL, ASP, XML programming and skills for creating Web pages.

Course projects in E-Business Systems

The course projects in E-Business Systems are tasks aimed at programming of Electronic Commerce (E-Commerce) Web sites with ASP.NET 3.5 (Active Server Pages .NET 3.5) and C#. They realize shop (firm) for E-Commerce. The set of functionalities, building a course project, is analogical to the same set of functionalities described in [5].

Phase one of the development of the project includes building a simple and usable product catalog for an appropriate database, that makes it easy for visitors to find the products they're looking for; the various departments of the shop (firm); the categories and subcategories of products and product attributes; searching the catalog for products by entering one or more keywords; catalog administration

for products, departments, categories and their attributes.

Phase two of development realizes the custom shopping cart (basket), which stores its data in the local database; editing and showing the contents of shopping cart; creating an order administration page; implementing a product recommendation system.

Phase three of development realizes building a customer accounts module so that customers can log in and retrieve their details every time they make an order; implementing the order pipeline that deals with credit card authorization, stock checking, shipping, email notification, and so on; adding full credit card transaction. The database should have at least 2 departments, at least 3 categories in department, at least 8 products in category.

5 More about the Educational Goals and Interrelations between Courses

There is a manifest continuity between the contents of the discussed three courses that has been extended and given meaning by the development of students' course projects.

From professional point of view, these courses are directed to mastering of some of the software design principles discussed in this section as a common objective.

Designing for Growth

This aspect is a crucial one because almost all users are more impressed with how a site looks and how easy it is to use than about which technologies and techniques are used behind the scenes or what operating system the Web server is running. If the site is slow, hard to use or remember, it just doesn't matter what rocket science was used to create it.

Unfortunately, this truth makes many inexperienced programmers underestimate the importance of the way the invisible part of the site – the code, the database, and so on – is implemented. The visual part of a site gets users interested to begin with, but its functionality makes them come back. A Web site can be implemented very quickly based on some initial requirements, but if not properly architected, it can become difficult, if not impossible, to change.

The discussed courses aim to acquire knowledge and experience in the development of online applications with scalable architecture. In a scalable system, in the ideal case, the ratio between the number of client requests and the hardware resources required to handle those requests is constant, even when the number of clients increases. An unscalable system can't deal with an increasing number of customers/users, no matter how many hardware resources are provided. Because we are optimistic about the number of customers, we must be sure that

the site will be able to deliver its functionality to a large number of users without throwing out errors or performing slowly.

Using a Three-Tier Architecture

The three-tier architecture is splitting an application's functionality unit into three logical tiers:

- the presentation tier;
- the business tier;
- the data tier.

The *presentation tier* contains the user interface elements of the site and includes all the logic that manages the interaction between the visitor and the client's business. This tier makes the whole site feel alive, and the way you design it is crucially important to the site's success. Since we suppose that the application is a Web site, its presentation tier is composed of dynamic Web pages.

The *business tier* (also called the middle tier) receives requests from the presentation tier and returns a result to this tier depending on the business logic it contains. Almost any event that happens in the presentation tier results in the business tier being called (except events that can be handled locally by the presentation tier, such as simple input data validation). Almost always, the business tier needs to call the data tier for information to respond to the presentation tier's request.

The *data tier* (sometimes referred to as the database tier) is responsible for storing the application's data and sending it to the business tier when requested. For the BalloonShop e-commerce site, one will need to store data about products (including their categories and their departments), users, shopping carts, and so on. Almost every client request finally results in the data tier being interrogated for information (except when previously retrieved data has been cached at the business tier or presentation tier levels).

These rules may look like limitations at first, but when utilizing an architecture, one needs to be consistent and obey its rules to reap the benefits. Sticking to the three-tier architecture ensures that the site remains easily updated or changed and adds a level of control over who or what can access the data. Therefore it is one of the principles followed in the discussed series of courses. Special attention has been paid to Microsoft technologies associated with every tier in the three-tier architecture (see Figure 2).

Interoperability

Interoperability in general is concerned with the capability of differing information systems to communicate. This communication may take various forms such as the transfer, exchange, transformation, mediation, migration or integration of information. If two or more systems are capable of communicating and exchanging data, they are displaying syntactic interoperability. Beyond the

ability of two (or more) computer systems to exchange information, semantic interoperability is the ability to automatically interpret the information exchanged meaningfully and accurately in order to produce useful results as defined by the end users of both systems. It involves:

- the processing of the shared information so that it is consistent with the intended meaning;
- the encoding of queries and presentation of information so that it conforms with the intended meaning regardless of the source of information.

Standardization and ontologies are indicated as most effective instruments for providing and maintaining interoperability in information systems.

Standardization may direct to the following favourable features of information systems:

- information can be immediately communicated (transferred, integrated, merged etc.) without transformation;
- information can be communicated without alteration;
- information can be kept in a single form;
- information of candidate sources can be enforced to be functionally complete for an envisaged integrated service.

Many different metadata schemes are being developed as standards across disciplines. As most popular metadata standards one should mention [6]:

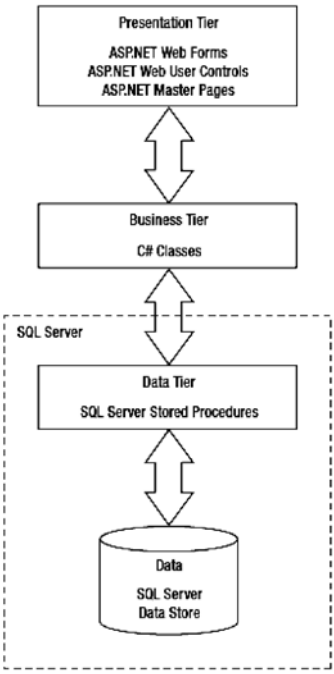


Fig. 2. Microsoft technologies and the three-tier architecture

- Dublin Core – an interoperable online metadata standard focused on networked resources;
- Encoded Archival Description (EAD) – a standard for encoding archival finding aids using XML in archival and manuscript repositories;
- Text Encoding Initiative (TEI) – a standard for the representation of texts in digital form;
- Z39.87 Data Dictionary – a technical metadata standard for a set of technical data elements required to manage digital image collections;
- Machine Readable Cataloging (MARC) – a set of standards for the representation and communication of bibliographic and related information in machine-readable form;
- Multimedia Content Description Interface (MPEG-7) – a ISO/IEC standard which specifies a set of descriptors to describe various types of multimedia information;
- Directory Interchange Format (DIF) – a descriptive and standardized format for exchanging information about scientific data sets.

Learning a considerable part of these standards is one of the educational goals of the discussed undergraduate courses.

6 Future Work

One of the well-accepted mechanisms for achieving semantic interoperability is the utilization of *ontologies*. According to the most popular definition, "an ontology is an explicit specification of a conceptualization". Metadata vocabularies and ontologies are considered as ways of providing semantic context in determining the relevance of resources.

The standards for metadata descriptions usually provide only general-purpose structures. Therefore, the utilization of Core and/or Upper Ontologies in capturing the semantics of the standards in combination with Domain Ontologies that extend them with domain knowledge, are systematic mechanisms for the extension and adaptation of standards [6, 7].

The notion of ontology as well as some popular languages and tools for building and using ontologies have been examined in another course included in the Information Systems curriculum – the one on Knowledge-based Systems. We are planning to extend the contents of the XML Programming course with knowledge about the W3C standards for description of information resources (in particular, simple ontologies) RDF and RDFS [8, 9] and their XML-based syntax. The curriculum of the course in E-Business Systems will be enriched with discussion of some popular ontologies applicable in E-Commerce.

Acknowledgments. This work has been funded by the Sofia University SRF within the "Methods and information technologies for ontology building, merging and using" Project, Contract No. 177/2012.

References

1. Harold, E.R.: XML 1.1 Bible, 3rd Edition. ISBN 0-7645-4986-3, Wiley Publishing, Inc., 2004.
2. Hunter, D. et al.: Beginning XML, 4th Edition (Programmer to Programmer). ISBN 978-0-470-11487-2, WROX Press, 2007.
3. ISO 8879:1986 Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML).
4. MacDonald, M.: Beginning ASP.NET 3.5 in C# 2008: From Novice to Professional, 2nd Edition. ISBN 978-1590598917, Apress, 2007.
5. Watson, K., Darie, C.: Beginning ASP.NET E-Commerce in C#: From Novice to Professional. ISBN 978-1430210740, Apress, 2009.
6. Nisheva-Pavlova, M.: Providing and Maintaining Interoperability in Digital Library Systems. Proceedings of the Fourth International Conference on Information Systems and Grid Technologies (Sofia, May 28-29, 2010), ISBN 978-954-07-3168-1, St. Kliment Ohridski Uiversity Press, 2010, pp. 200-208.
7. Nisheva-Pavlova, M.: Mapping and Merging Domain Ontologies in Digital Library Systems. Proceedings of the Fifth International Conference on Information Systems and Grid Technologies (Sofia, May 27-28, 2011), ISSN 1314-4855, Sofia, St. Kliment Ohridski University Press, 2011, pp. 107-113.
8. Champin, P.: RDF Tutorial. http://bat710.univ-lyon1.fr/~champin/rdf-tutorial/ (visited on May 25, 2012).
9. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, 10 February 2004. http://www.w3.org/TR/rdf-schema/ (visited on May 25, 2012).

Web-based System for Teamwork Evaluation

Radoslava Hristova[1], Goran Goranov[2], Elena Hinova[3]

Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5 James Baucher, 1164 Sofia, Bulgaria, radoslava@fmi.uni-sofia.bg
[2] Faculty of Electrical Engineering and Electronics, Technical University of Gabrovo,
4 Hadji Dimitar str., 5300 Gabrovo, Bulgaria, g_goranov@tugab.bg
[3] AKTA Medika CEO, 60 Nikola Petkov str., 5400 Sevlievo, Bulgaria, ehinova@aktamedika.org

Abstract. In this article we present a web-based system for competence evaluation, based on the 360 degree feedback method. The system is developed to meet the requirements of a medical center and provides functionality for assessment of the medical center's employees. The benefits of the presented system are improvement of the time for data processing and graphical representation of the collected data. The system helps manages to do the right evaluation of the employees, based on the presented assessment method and reduces the errors of calculation for the evaluation.

Keywords: relational databases, competences, assessment, 360 degree feadback

Introduction

The 360 degree feedback method is a combination of two contemporary methods for assessment [1] – the self assessment method and the peer assessment method. As a combination, the 360 degree feedback method combines the result from evaluation of the two assessment methods:

- The self assessment of the person and
- The assessment of the people who work with this person.

Oftent the results of the self assessment method are not so reliable. The reason is that the persons are not inclined to be objective. They either underestimate themselves or overestimate themselves. Combination of the self assessment with the peer assessment can give more exact evaluation. The features of the 360 degree feedback method [2] are:

- measure of person's behaviors and competencies;
- feedback on how other people perceive the evaluated person;
- measure of person's skills such as listening, planning, and goal-setting;

The 360 degree feedback method focuses on behaviors and competencies more than on basic skills, job requirements, and performance objectives. For these

reasons, the 360 degree feedback method is more often used then other methods in the organizations (business organizations or universities) for evaluation.

For example – companies. The companies usually have three-level hierarchy, which typically include the employee's manager, peers, and direct reports. The 360 degree feedback method includes self assessment of the employee and the anonymous feedback from the people who work with him – his manager, peers and his subordinates (Figure 1).

Fig. 1. The 360 degree feedback method three-level hierarchy assessment

The universities usually have two-level hierarchy, which tipically includes the student's teacher and the peers. In this case the assessment method is known as 180 degree feedback method.

The evaluation process consists of feedback forms, which have to be filled by raters, for example the manager and the peers. The feedback forms include questions, which are measured on a rating scale. The person who receives feedback also fills out a self-rating survey that includes the same survey questions that others receive in their forms. At the end of the process all the answers have to be evaluated, generalized and presented as aggregated results.

If the process is processed by a human, two major disadvantages can be noted. The assessment has to be confidental. In the case of human work, this confidentiality can be violated. Second the data have to be processed by the human and presented graphically. This usually requires additional time for data representation.

The whole 360 feadback process can be automated. This is where the 360 feedback systems take place. There are different solutions of the 360 feedback systems. Example ones are provided by [2], [3] and [4]. The advantages of the 360 degree feedback systems to the traditional human work can be generalized as follows:

- The 360 degree feedback systems are used for teamwork evalution;

- The feedback process automation gives people an opportunity to provide anonymous feedback to a coworker that they might otherwise be uncomfortable to give;
- The system automatically tabulates the results and presents them in a format that helps the result to be analysed.

In this article we present a web-based system for competence evaluation, based on the 360 degree method. The system is developed to meet the requirements of a medical center and provides functionality for evaluation of the medical center's employees. We present them in the next sections.

Functional and supplementary requirements

The web-based system, which we discuss in this article is developed to meet the requirements of the medical center - AKTA Medika [5] and to provide functionality for evaluation of the medical center's employees. The managers of the medical facility put the following main requiremnts to the system:

- Each employee have to be registered into the system;
- The system has to provide administrative panel fo user creation and updates;
- For each employee form feedback have to be full-filled;
- The raters are the employee himself, his manager, his peers and his subordinates;
- According to the user's role each evaluation has different weight – the weights of the manager's, the subordinates' and the peers' evaluations are heavier then the weight of the self-evaluation of the employee;
- The system has to provide features for generalization and aggregation of the feedback results;
- The system has to provide features for creating easy reports. The desirable results are graphically presented data. The reports have to include comperative analysis between the evaluated employees and their peers. The reports also have to include data for every evaluated employee, grouped by competencies and directions for future growth.
- The systems have to be extnedable. It has to provide features for creating new test evaluations, for adding new competences etc.

System design and implementation

Based on the above requirements a classical three-layered system was developed (Figure 2). On the first layer is used relational database for storing data and representation. On the second layer is used web server (Apache HTTP Server 2.0), where the main logic of the system was implemented. On the third layer

is used standart web client for access to the system. The system is web-based, thus it is accessible from everywhere through internet browser. This makes the system usable and adaptable for any other cases, for example other companies or universities, which need a 360 degree feedback solution.

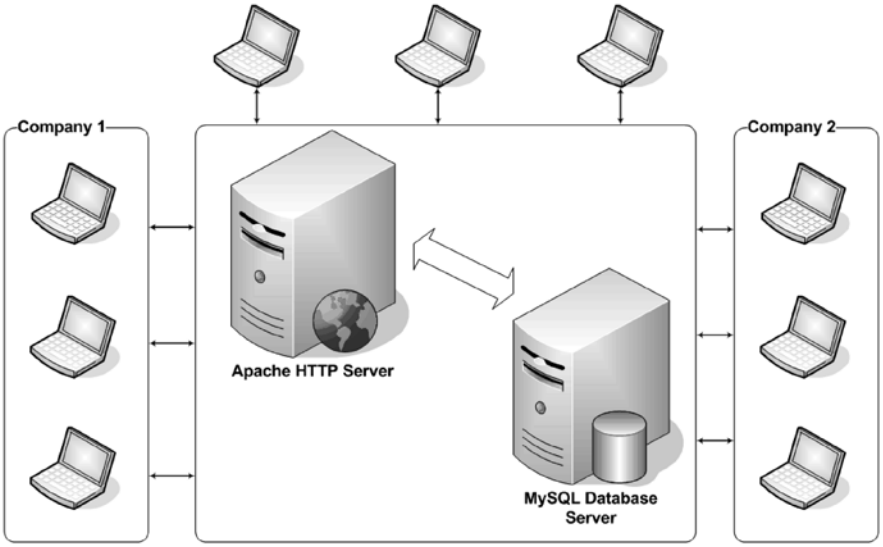


Fig. 2. System design and implementation

As relational database system is used MySQL 5.1 Database. On (Figure 3) is presented the database model with the basic tables and the relationship between them. The database stores information for all the employees, evaluations, the different competencies and items related with them. Based on the collected data, various reports can be created.

A web-based interface for access to the system was developed [6]. The developed application is based on PHP 5.0. It provides two type of access to the system administrative access and user access. The administrative access allows the following actions:

- Create or update users;
- Create or update competences;
- Create or update items;
- Create or update tests;

Example usecase how the system can be used from the administrative panel is as follows:

1. The administrator create user;
2. The administrator create test and set the deadline when the test finishes; Main elements of the test are the test's name and the name of the person

who will be evaluated;

3. The administrator chooses the employees who will participate into the evaluation and who will rate the person;
4. The administrator chooses from existing items and competences, and adds them as active for the current evaluation.

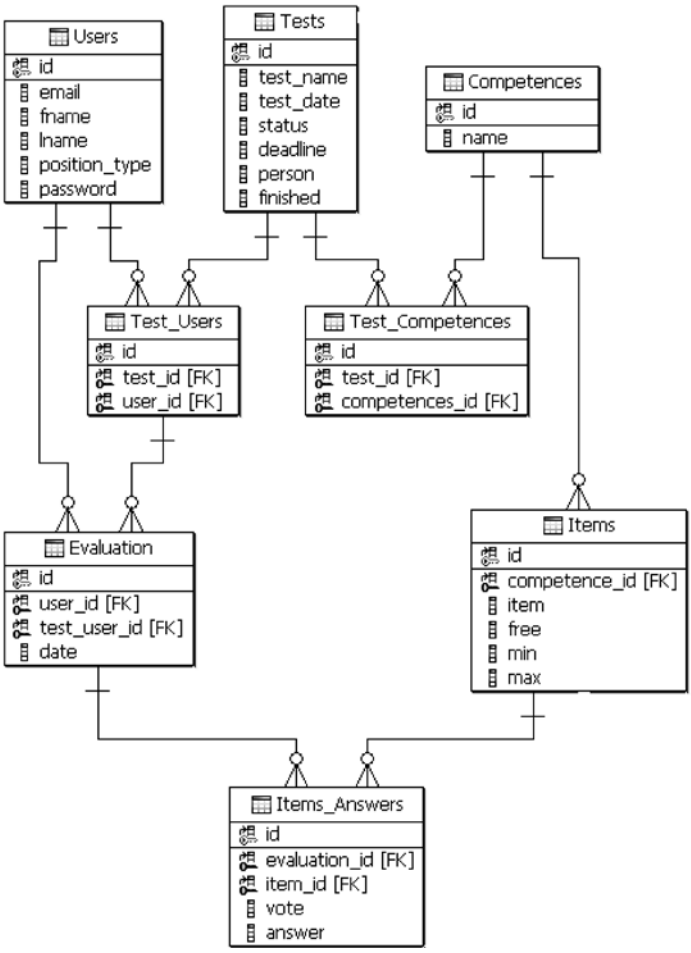


Fig. 3. Database relational model diagram

Example usecase how the system can be used from the user panel is as follows:

1. The use logins to the system;
2. The user chooses test;
3. The user fills the forms and evaluates the person for whom the test is

created;

4. The user can filled the test just once;
5. After the user finish with filling the test, the test is closed and is no longer available.

All reports and statistics can be done from the administrative panel. On (Figure 4) are shown system's menu items that the administrator can access through the system.

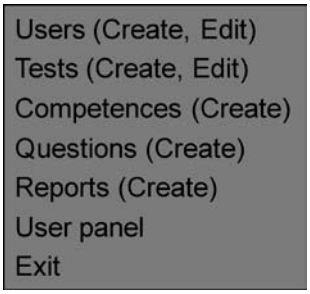


Fig. 4. System's menu items

The aggregated evaluation can be visualised as graphics. The evaluation results are grouped by competences, grouped by directions for future growth or displayed into details. The results are rated according to weight scale: excellent, competent and satisfaction. The report can be printed.

Conclusions

In this article we presented a web-based system for teamwork evaluation, based on 360 degree feedback method. The benefits of the presented system are improvement of the time for data processing and graphical representation of collected data. The system helps manages to do the right assessment of the employees, based on the evaluation method and reduces the errors of weights calculation for the evaluation.

References

1. Kirkova - Bogdanova, A., Modern assessment methods (Master Thesis), Sofia University "St. Kliment Ohridski", Faculty of Mathematics and Informatics, Department of Information Technologies, http://research.uni-sofia.bg/bitstream/handle/10506/194/Abstract%20Eng.pdf
2. 360 Degree Feedback Solutions, http://www.custominsight.com/360-degree-feedback/what-is-360-degree-feedback.asp
3. 360 Degree Feedback from Team Builders Plus, http://www.360-degreefeedback.com
4. DecisionWise Services, http://www.decision-wise.com/what-is-360-degree-feedback.html
5. AKTA Medika, http://www.aktamedika.org/
6. 360 Degree Feedback Web-based platform realisation, http://360-bg.info/

System Analysis of Information Systems for Local Economic Development Modeling – Case Study for the Region Pelagonia in R. of Macedonia

Snezana Savoska[1] , Branko Dimeskil[2],

[1,2] Faculty of administration and Information systems Management, University „St.Kliment Ohridski“ – Bitola,
Bitolska bb,
7000 Bitola, R.of Macedonia,
savoskasnezana@gmail.com, branko_dim@yahoo.com

Abstract. The system analysis is the necessary activity in designing information systems (IS), especially in creating complex IS which have to satisfy a wide pallet of users' demands. Installing the IS without expert's planning and leading can lead to the huge users' dissatisfaction, and maybe non - usage of system which often the consequent system do not work. This is especially emphasized when we talk about web based IS which demand a strong defined access rules as well as accurate data update procedures. In this paper is made a system analysis and design of IS for local economic development (LED). The problem of LED itself is very important because of the decentralization process that happens in R.of Macedonia in recent year as well as the global crises and the necessity of employment increasing. As an important factor we mention a need of increasing the usage of IS, especially when is concern of the issues that help for the young people's position. Analysis of the need of IS for LED's support is made on the couple of present local governments' (LG) web sites in R.of Macedonia as well as the interviews and a questionnaire of the LER's responsible in the LG and potential users of this kind of information. The results of this survey are decanting in analysis of the information needs as well as the LED's support System's proposition. We are using the structural analysis and logical design of IS as the working' methodology. For this purpose, a series of systematic undertaken processes were used. These processes, we think, that will enhance the information and usage of computer's IS in function of business climate and business community's better information. The proposed model for LED's support IS which have to cover the users' demands will be made with creating a redundant databases, loaded whit trigger procedures and intelligent agents.

Keywords: System analysis, Logical design, DFD, ER diagrams, Z-specification, Local economic development.

1 Introduction

The system analysis is an activity in analysis and IS design that includes a scientific approach for information gaining for "it is" IS as well as collecting users' information demands. It is particularly necessary in creating of wider IS or

the complex ones which have to meet the demands of the biggest users' groups and to satisfy a wide range of users' demands. Installing this kind of IS without the expert's planning and guides lead to huge users' dissatisfaction, causes the system do not satisfy the users' needs and in final – the dysfunctions of IS (Langar,2008). The system analysis of the IS which have to support LED is necessary activity because this concept is extremely complex which demands systematic approach in defining the information needs.

LED is a term with that is increasingly exploited in this region, especially with decentralization in R. of Macedonia and separating local from the central government. Globally, this concept of localization is implemented in different manner and can be found good or less good examples of LG's concept implementation. But, the fact is that in the globalization era, the role of LED for municipalities' development is undoubtedly huge, especially in gaining information and conditions for concrete support to the business community as well as all individuals which won't to start a new business. One of the most important LED's supports will be development of inventive environment which will contribute for organizational culture changes and also creating conditions which will enable the young people to start new businesses and prevent their leaving from R. of Macedonia, which is evident n the last years. One positive sparkle in this situation is the IT's global development which enable couple of young people to work – are employed in the global companies via internet connections (free lenders). But, this is not only way for IT to help in this manner. With usage of IT, there is strong influence of the level of transparency and information knowledge which is necessary for the business.

For those reasons, the primary goal in the paper will be to create system analysis and logical design of IS for support of LED. With this system analysis the "it is" IS will be detect the gap between the "it is" and desired "to be" IS. Also, the need of information for more intensive LED and data sources disposal in that moment will be detected. With system design the system's proposition for IS will be modeled and it has to meet the needs of potential users of LED's information. For this objective, it is necessary to analyze the work of LG's and governments' administrative institutions which have the common point with business sector. Also, there is necessary to analyze data which is necessary for LED's promoting in the LG in R. of Macedonia, but obtained from the potential survey about the use of LED's data.

2 The Preparation for System Analysis and Logical Design

Planning the usage of information communication technology in the Local government first of all means satisfying the needs of transparent information which LGs are obligated to provide (Valacich, 2006). In almost all local governments in R .of Macedonia the transparency is achieved with Content management

web sites[1] where located information is accessible for public usage. In this way, more or less, LGs provide transparent information that have to be transparent and for public usage. But this information is huge and unstructured, difficult for researching and without defined dictionary for the means, concepts and notations. However, almost all of them possess information with transparent character (contacts with institutions, important telephones, on-line reading of local informers, organizational structure, ongoing projects and the other). Nevertheless, we mean that these sites do not provide real useful information for the business community to find the desired information on the sites if they haven't enough IT skills. Also, many of them are not ordinary updated (some of them are not updated many years), they are static (without the user's interaction, except the e-mailing possibility to send to contact person, which usually do not provide the answer) and do not possess a part for inventiveness development which is very important for LED's development.

The LED employees can be contacted beside e-mail also by the blogs, which are the popular tools for communication and collaboration as well as the other forms of Wiki-logs (blogs) or by the social frameworks as Facebook, Twiter and the similar ones. Some of LG already have their profiles but, usually they provide for information to same or similar policy opinion members. Ordinary, the information exchange between the business community and LG unit's produce outputs, but they are not provided at all almost in all LG's in the R .of Macedonia.

3 System Analysis of the Available IS for LED's support

To obtain more accurate apprehension for overall problem that is the issue of our analysis, first we have to answer the following research questions:

1. Which IT policy exist at the local level in the LGs in R .of Macedonia or strategies for IT development in order to achieve a quality infrastructure which have to help business sector and to foster LED in LG?

2. Are some of these policies successful implemented in the LG? If they are, in which way can be improved in order to gain more qualitative and more accurate on-time information for the business community?

3. If there are not defined policies for LED' support, how they have to be defined? What kind of IS has to be created, with which procedures for LED business community support?

3.1 Unobtrusive Methods for Information Gathering

The analysis with unobtrusive method is done on the base of LG's web sites analysis. In this case we get information about the LG's policy for IT support

[1] Usually the hardware and software in LG are obtain as donations from donors and are installed without previously created strategy for integrated IT development of LG, only for providing transparency for the Macedonian citizens;

for LED. Although in many of the LED's strategies we found relatively accurate defined information which is available for LED, yet they don't have a real support, they do not contain the real need data neither links to that data[2] from LED strategy. Somewhere data wasn't updated couple of years. Also, we met data for international collaboration and their links which suite call for project funding (fund raising). Most of the LG in R. of Macedonia already consider about creating Information communication technology (ICT) strategy in which frame they will provide a part for LED support. Some of LGs already have the ICT strategies which are accessible from the LG's web sites. But, the fact is that there is only a declarative support for these strategies in the LG not only because they contain the prerequisites and the needed data which LG have to provide, but they contain neither the real databases nor the links to the specific institutions' links where the data can be found. The researchers were made on the analysis of twelve web sites of local governments in R .of Macedonia.

In almost all LG in R .of Macedonia exists sectors for local economic development. Usually they have announced the main information which is in the scope from defining of their competency (LED planning, creating strategies and priorities, law obligations etc.), to the contact person in this sector and links to the relevant institutions, agencies, NGO and other factors. Besides that, the information for the facts about LG are published as well as the information which has to be published announced. The data that they deem important for LED include databases on the situation in LG, analysis of business climate, opportunities, strengths, weaknesses and threats to business. Although these data can be significantly important for the business community, they do not transparently publish, but should be looking after advertisement that exists on their web pages. However, when it comes to strategic planning of the use of ICT in the LED, we can say that there are significant results.

3.2 The Detected Information Need for LED

After the phase of using unobtrusive methods of obtaining information about LED, with a population survey of the business community, consumers and potential users of the web site to support the LED, the most important data that businesses need to get the IS for this purpose are:

- Data on existing business subjects in the municipality (LG);
- Data on supply and demand of labor in local government and social map of LG;
- Data available for business locations and available utility infrastructure;
- Data for road, telecommunication and logistics structure;
- Data on economic events, fairs, other events;

[2] For example, http://www.opstinagpetrov.gov.mk/index.php?option=com_content&view=article&id=2372&Itemid=100

- Data required documentation for registration of a company, tax obligations, tax benefits and other legislation;
- Donor organizations, banks and credit opportunities;
- Benefits of International Business Ventures;
- Help with creating a business plan and other start-up activities;
- Data to support innovation and development activities;
- Data for needed goods and services in the municipality;

In the phase of using interactive methods for exploring the information requirements we have incorporated these questions and tried to get answers from those interviewed.

3.3 Interactive Methods Used in System Analysis for LED Support

By using structured interviews (two) and questionnaires (five) with employees in LG and interviews with potential users of information on LED (10 questionnaires), we tried to get the required information directly from the top people in charge of the LED in municipalities and the latter directly from businesses – potential users of IS for the LED. We put 16 questions for specific IT support LED in municipalities and 10 issues of customer information. From the analysis of responses to the first interview and asked groups we could get their thoughts for the need for LED IT support and also for the current support which they give to the business sector. The general conclusion is that local government employees still have a "bureaucratic view" of problems that are not put into function (operation) of the LED, as predicted EU policies for transforming of Public administration (PA). They generally believe that what we have on web pages to municipalities and ministries, government agencies are sufficient for the business sector. Also it is considered that the Chamber, if the business entity is subscribed, it gives enough information about investment opportunities, calls for projects, etc. (Davis, 2009). Also, electronic documents are available on the web pages and can be downloaded from there.

As to the second group questioned, it is important that they feel that they haven't enough knowledge to themselves to obtain necessary information, especially when dealing with subtle issues such as finding information for the new businesses, relationships with foreign partners, donors or partnerships with foreign companies. Considering that the required information is better placed in one place and has links to the necessary institutions where to seek information about themselves. The important thing about them is that they haven't sufficient funds to pay for information (as in the case of membership in the Chamber, where there is already some information on possible business partnerships, funding opportunities, etc.) and expect it to get the LED offices in LG. Prefer web-based IS that will bring all the information on the desktop of potential users, and this information will be accurate, timely and successfully managed from LED in LG.

4. Analysis of Obtained Results and Defining the Information Needs for the LED's IS Users

If we analyze the needs of potential users of IS for LED support for transparent information it is necessary to create a web oriented IS which will merge the disparate data through a common database that would be redundant for data but consistent for the LED's needs. It would be loaded with triggers – some kind of intelligent agents that will provide timely data (using the trigger procedures) and predefined data sources. This software application would implement the necessary links to sites where the data are extracted and the sites of donors, banks, governmental and nongovernmental agencies and organizations and all information what can help LG's LED. Contacts with staff in LED Department of LG may, except e-mail to be done with the help of blogs, and other types of collaboration such as discussion groups, expert groups etc. Although some information is not completely public, it is necessary to find a way to see them by interested or be willing to form and shape that will not endanger the Low of Protection of personal data[3] and will be in accordance with Law on Free Access to Public Information[4]. Although the LG have a piece of information that have to be included in the databases, they are not prepared adequately for use, but in Office documents and should additionally be transformed into a format suitable for placing in the databases. For support of inventive capacity of the business community, it is possible to use the collaboration tools like blogs, discussion groups and expert groups where besides LED experts will participates the universities – i.e. teachers and students which will be engaged in research on market opportunities and the other critical areas.

5. Preparing the Systems Proposal – The Logical Structure of the System Proposal and DFD

The logical design of IS for LED support should give those information and flows that are necessary for everyday users. It will be made using the tools: DFD – Data flow diagram and ER (Entity Relationship) diagram. The first one is a graphical representation of processes and connections between them as well as system inputs, system outputs and databases, while the second one defines the entities and links between them which are the basis for further definition of databases. Based on the analysis, we defined DFD diagram of proposed IS for LED support, shown in Figure 1. Each DFD process diagram can be put to single forms as end screens – single processes. It is important that the complex DFD diagrams can be partitioned in accordance to the software engineering principle with purpose to be more manageable. Therefore, in the Figure 1 is shown a part of IS for LED support as

[3] http://www.uvmk.gov.mk/files/zakoni/ZZLP_Precisten%20tekst.pdf

[4] http://www.stat.gov.mk/pdf/SlobodenPristapDoInformacii/zakon_za_informacii.pdf

decomposed DFD diagram of the process. During decomposition, the principles of conservation and balancing the DFD diagrams were kept (Kendall, 2007). For each process, depending on the process' complexity, you can write logic of documenting and analyzing the processes and tools with structured English, decision tables and decision trees (FIS, 2008).

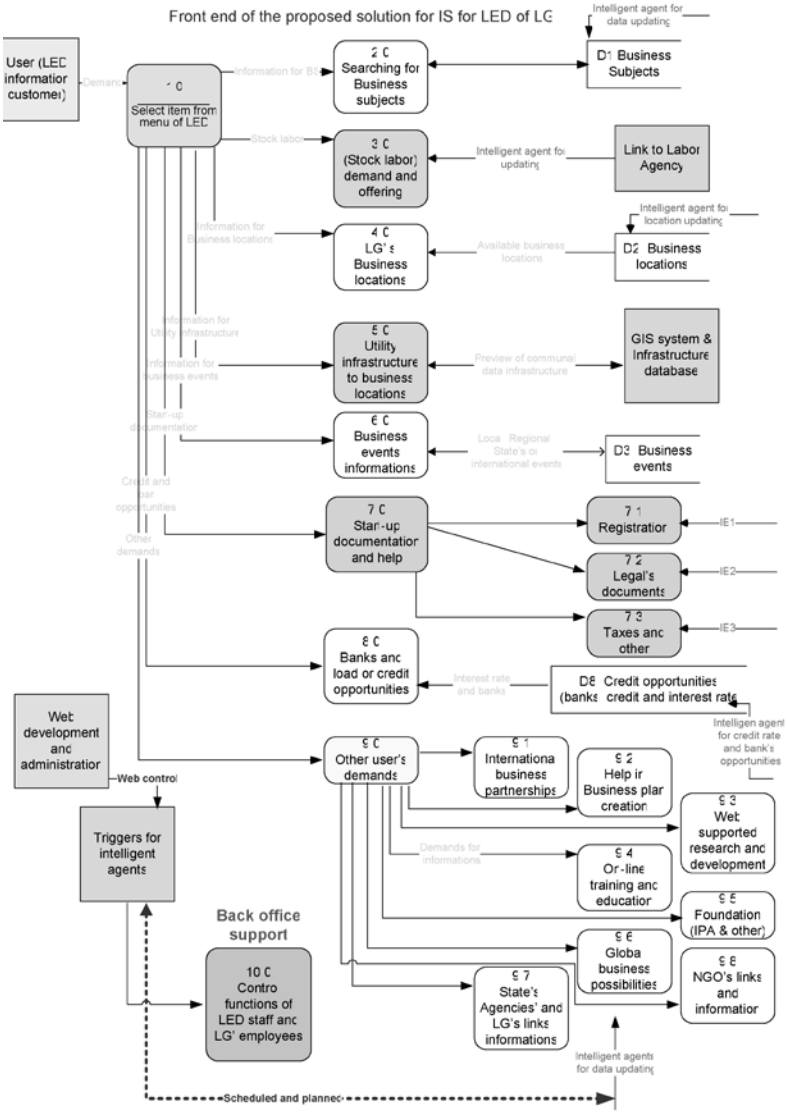


Fig. 1. The proposed DFD diagram for IS for LED support

6. ER diagrams for IS for LED support

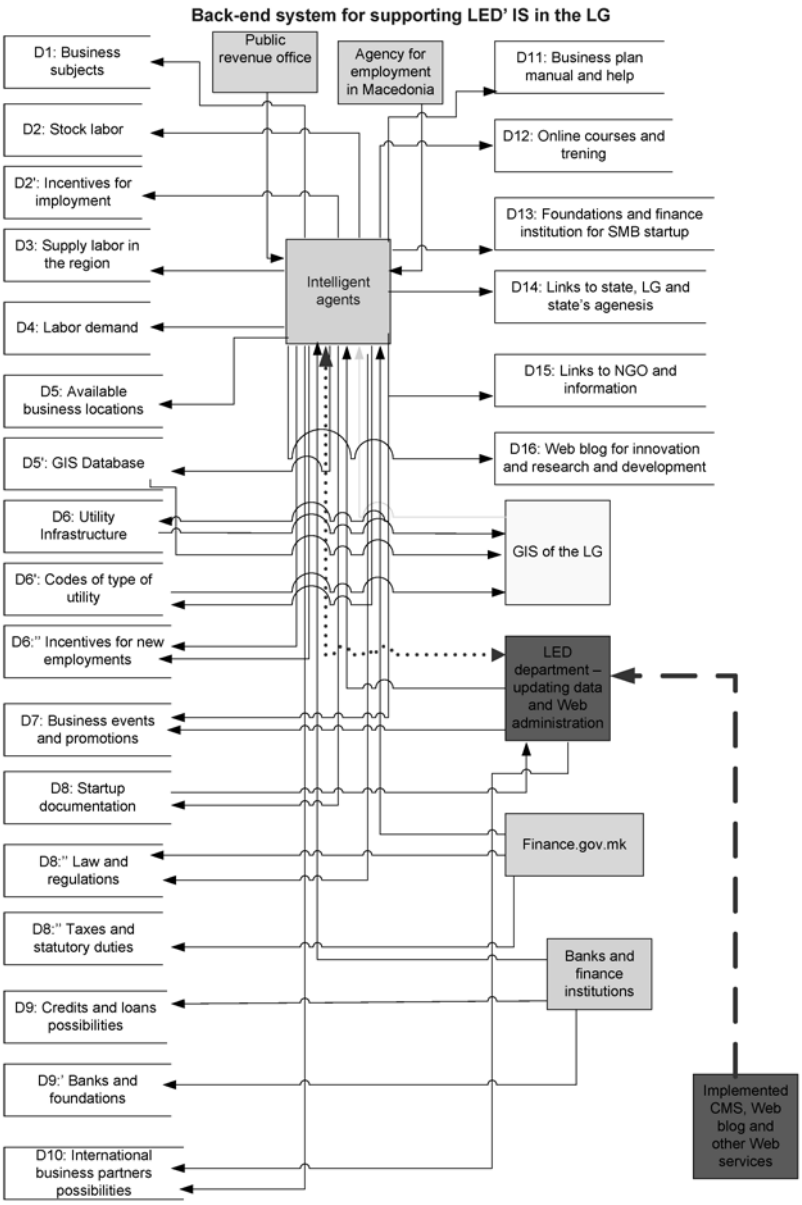


Fig. 2 The structure of the database from the proposed solution for IS for LED support

The definition of entities and relations between them is a part of the physical design of the IS, so we define ER diagrams of some of the entities – objects related to a process of logical DFD diagram was first decant in the physical DFD diagram. Because of the limited space in this paper, only the part of the structure of the ER diagram is shown. It is represented by databases – as groups, parts of the snow-flake structure of databases. Defining the logical diagram is just one step toward the design of the system. While he defines business processes which have to be supported by IS and close to understands for non-IT staff, system analyst should prepare a proposal for system developers for IS for LED support. They have to prepare in detail the physical structure of IS based on physical DFD diagram and ER diagrams. They should be prepared on the logical diagram base and also the business logic which have to be supported by IS for LED support. Example of a physical diagram of the processes is presented in Figure 3.

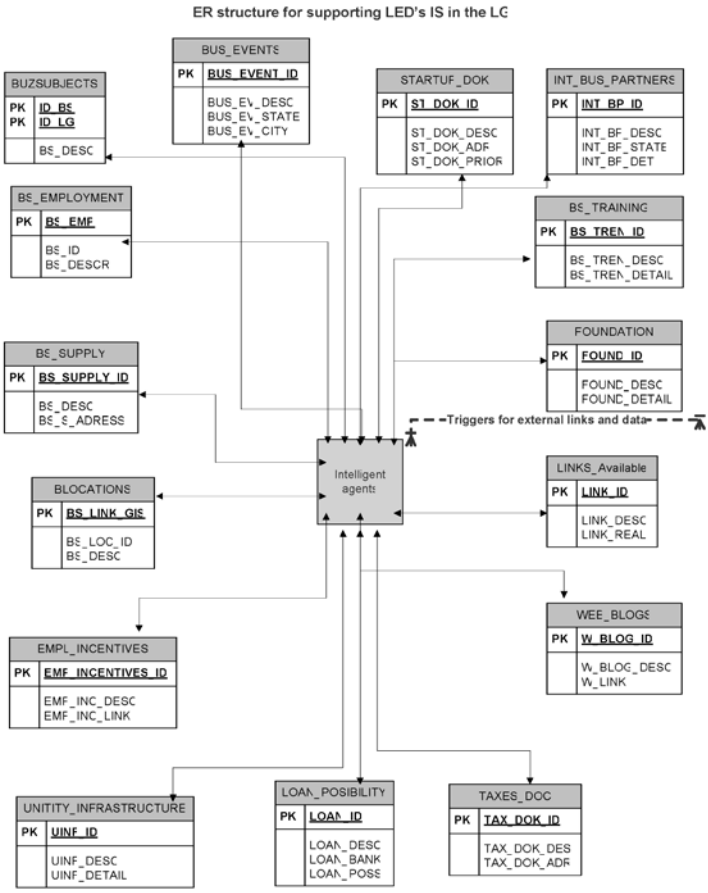


Fig. 3. ER diagram for IS for LED support

7. Integration of logical and physical diagram and ER diagram and systems proposal

When the logical and physical DFD diagrams, ER diagrams are defined and the logic that specifies processes DFD diagrams is also defined, you can start by creating a system proposal for IS for LED support. Despite detailed diagrams, it should include cover, title page, and contents of the proposal, a summary for the donors of the project, notes and proper documentation and also detailed results of system studies, system alternatives and recommendations of system analysts (Kendall, 2007). It should be described in the summary and followed by accessories which will contain all analyses, forms, diagrams and working documents. In this documentation must be found also the value of the investment and to evaluate efforts for system analysis and design of IS (Oestereich,2001). The physical system, as it will use the users should be developed in end-users screens that need to intercept the detected information needs (Bowman, 2004) and should translate into concrete solution. In the testing phase, it is necessary to compare the requirements with the achieved outcomes and to evaluate the new IS also achieved customer satisfaction, followed by implementation of the system proposal.

To better explain the proposed system, it is good to make the relational model of data for the system. Relational model presented using UML diagrams would help in the qualitative analysis of the system. Furthermore, we can use Z-specification for development of UML model at higher level of abstraction, as it is explained in [9,10].

8 Conclusion

The creation of IS for LED support is the present challenge, because of the need arising from decentralization and also because of complexity which is not only a challenge for developers, but also for members of the LED community in Macedonia. The successful design of the project largely depends on proper system analysis as well as properly established logical model of IS. Certainly the greatest benefits from the implementation of the project will get users of IS to support LED and the business community, and thus the municipalities and the state. However, logical design, physical design, development model and the programming of web-based solution means a successful implementation. It will certainly depend on a number of other factors such as dissemination of the use of IS, conducted training for the unemployed and small and medium sized companies for its use and engagement of employees in local government responsible for LED. In this phase NGOs can be included primarily in promotion of usage of the IS and indirectly in training, which should be allowed from the local government'

site. Therefore our future research will be focused on organizational aspects of implementation and the presentational interface and how data will be provided and will result in the databases of IS for LED support.

References

1. Langer A. M., Analysis and Design of Information Systems, Springer, 2008
2. Kendall E.K. System analysis and design, Prentice Hall, 2007
3. Valacich, George, Hoffer, Essentials of system analysis and design, Prentice Hall, 2006
4. Oestereich B., Developping software with UML, object-oriented analysis and design in practice, Addison-Wesley, 2001
5. Booch G., Object oriented analysis and design with application, Addison-Weslet, 1998
6. Bowman K., System analysis, a beginners' guide, Palgrave, 2004
7. http://www.prilep.gov.mk/index.php?option=com_content&task=view&id=915&Itemid=1, април, 2012, Општина Прилеп, 2011-2012
8. Davis B., 97 Things every project manager should know, O'Reilly, 2009
9. Dimitrov, V.: Formal Specification of Relational Model of Data in Z-Notation, Proc. of XXXIX Conf. of UMB, 178-183 (2010)
10. Dimitrov, V.: "Relationship" Specified in Z-Notation, Physics of Elementary Particles and Atomic Nuclei, Letters, Vol. 8, No. 4(167), 655—663 (2011)

A Comparison of Change Management Implementation in ITIL, CMMI and Project Management

Kalinka Kaloyanova[1], Emanuela Mitreva[1],

[1]Faculty of Mathematics and Informatics, Sofia University, 5 James Bourchier blvd., 1164, Sofia, Bulgaria

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

Abstract. The paper focuses on the Change Management process. The scope of the process, its main goals, roles, politics and procedures are discussed. Three methodologies that apply Change Management are compared – ITIL (Information technology Infrastructure Library), CMMI (Capability Maturity Model Integration) and PM (Project Management). A method is presented that enables to compare the implementing of the process into different methodologies. The paper contributes to the understanding of the adoption of Change Management in various IT areas.

Keywords: ITIL, CMMI, Project Management, Change Management

1. Introduction

All areas of today's business world change very rapidly. This is not only requires the organizations to manage the effectively and efficiently delivery of the changes but also is a key factor for them in order to be competitive. The ability to react to change requests and to control the change process is an important part of the work of every IT department.

The Change Management is an organizational process that supports changes accepting in the business environment. Change Management helps organizations to understand and to work for minimizing risks of changes to the IT environment. The main focus of the Change Management process is to manage and control the changes and to formalize the call flow, the procedures of change and the ways to measure the success of the application of a change. Formal procedures and activities make the risks more predictable and clear. This is the key goal for the Change Management process - to ensure that all concerned parties' requirements for quality, functionality and effectiveness are met.

The Change Management concerns different IT aspects from software products and services development and maintenance [9], [6] to controlling changes in the complex IT environment [12], [13]. All organizations would have well-documented Change Management practices, strong procedures how to follow them and tools that support the process [11].

In this paper the changes and their aspects are discussed in different frameworks from different point of view. The implementations of the Change Management process in three methodologies are compared: ITIL (Information Technology Infrastructure Library) [13], CMMI (Capability Maturity Model Integration) [18] and PM (Project Management) [5]. Our study also reveals best practices of managing changes in these different areas in order to provide a deeper understanding of the process and its adoption.

2. Applying Change Management in Different Methodologies

ITIL

ITIL (IT Infrastructure Library) is a framework of good practices, which include more than thirty processes that help companies to organize more efficiently and effectively the work of their IT departments. In ITIL v3 the service lifecycle is formed by five phases and all processes are divided into these five main phases: Strategy, Design, Transition, Operation, and Continual service Improvement [12].

Change requests could be generated from different parts of the organizations for a different reason. Although Change Management formally is a part of the Transition phase, changes can be initiated by any of the processes and any of them can result into inputs for the Change Management process.

The purpose of Change Management in ITIL is to make sure that standardized methods and procedures are used for efficient and prompt handling of all changes. Change Management in ITIL provide a well defined, measurable and consistent process for all service changes.

In order to ensure that all changes will be applied in a controlled and managed way, that will cause as little disruptions of the services as possible [12] ITIL introduce a number of definitions, politics, procedures, activities, and roles that help all changes to be recorded, evaluated, planned, tested and implemented.

For these reasons organizations that are starting on the path to introduce a formal Change Management process often come across the best practices described in the Information Technology Infrastructure Library.

CMMI

Another area where Change Management plays important role is Capability Maturity Model Integration (CMMI). The main goal of CMMI is to help organizations to improve their performance. Developed to provide guidance for developing processes that meet the business goals CMMI may also be used as a framework for appraising the process maturity of the organizations.

CMMI exists in two representations: continuous and staged. The staged representation defines five levels of maturity: Initial (processes are unpredictable and uncontrolled), Managed (processes are characterized and reactive measures are undertaken), Defined (processes are characterized and proactive measures are undertaken), Quantitatively managed (processes are measured and controlled), Optimizing (focus improvement). As staged representation is designed to provide the standard guidances for improvement, the continuous representation is designed to allow the users to focus on specific processes. CMMI also supports as an area of interest Product and service development - CMMI for Services (CMMI-SVC).

In the staged CMMI the Change Management process is formally a part of the Configuration Management process, which is in the maturity Level 2, and is one of the most important processes in CMMI. Its purpose is to establish and maintain the integrity of work products.

Project Management

Project management is the discipline of planning, organizing, securing, managing, leading, and controlling resources to achieve specific goals [14]. The description of the Change Management process in PM will be based upon PMBOK (Project Management Book of Knowledge). PMBOK recognises 5 basic process groups (Initiating, Planning, Executing, Monitoring and Controlling, Closing) and nine knowledge areas typical of almost all projects. The Change Management process is not part of any of the knowledge areas, but change requests and changes can be initiated by any of the processes.

PMBOK describes Change Management as a process, tools and techniques to manage the people-side of the change to achieve the required business outcome. Change Management incorporates the organizational tools that can be utilized to help individuals make successful personal transitions resulting in the adoption and realization of change.

3. A Comparison of Change Management in ITIL, CMMI and Project Management

When comparing Change Management process in different areas a set of various characteristics could be investigated. The main obstacle here is to find the right way of a valuable comparison.

From the above discussed methodologies ITIL presents the best process description. First, ITIL provides a brief description of the process, providing information about the main goals and the scope of the process. The scope of each process can be intertwined with other processes or it can vary in the different

methodologies.

Other very important points of a process are the politics, which defines how the process should be applied and the main call flow of the application of the process, which on its turn defines the main activities and roles. For each step of the call flow, ITIL is providing the corresponding activity and respectively a person, which is responsible for the activity.

The next significant topics are the description of the inputs/outputs of the Change Management process and metrics. The metrics play essential role in ITIL because they offer a way to measure the success rate of a process and ways of improvement for the future implementations of the process.

Based on ITIL well-defined structure of describing a process we establish a set of characteristics that describe the main points of the CM process. The context of the our comparative analysis is specified by the following elements:

CM comparative analysis = {Goals and Scope; Politics, Main Call Flow (Activities) and Roles; Input/Output and Metrics}

The next section of the article presents the similarities and differences between the application of the Change Management process in ITIL, CMMI and Project Management.

Goals, Scope

The ITIL process description starts with description of the goals (what the process should achieve) and the scope (what are the boundaries of the process and what is and isn't included in its main responsibility for the final delivery of the service).

Table 1. Scope and goals of the Change Management process in ITIL, CMMI and PM

	ITIL	**Project Management**	**CMMI**
Goals	- approving, assessing, and applying the changes without disturbance of the service; - risk assessment; - optimizing the time during which the service is down; - application of the changes in a way to minimize the downtime; - change approval of each change which is applied on the system. [13]	- identifying, assessment and control of the changes [15]; - approval of the changes which should be applied [19]; - risk assessment [19]; - optimizing the time during which the service is down [3]; - application of the changes in a way to minimize the downtime [17]; - change approval of each change which is applied on the system [16].	- identifying, assessing and controlling of the changes; - approving the changes, which should be applied; - optimizing the time, during which the service is down. [10]

Scope	- prioritizing and complying with the customer's requirement for change; - minimizing failed changes and disruption of the services; - applying the changes on time; - risk assessment; - minimizing downtime of the services by minimizing high level emergency changes; - process monitoring and identifying ways for improvement. [13]	- identifying who will pay for the changes; - who will be controlling the changes; - identifying of rules for managing change requests; - change schedule; - cost assessment; - impact assessment; - risk assessment; - monitoring and controlling the changes. [7]	- approval/rejection of the changes; - impact assessment; - monitoring and closing of the change; - tests; - regression tests. [10]

Table 1 presents the comparison of the main goals and scope in Change Management. What is obvious is that the goals of the process in all three methodologies are similar and therefore it will result into the same call flows and as we already mentioned – almost the same activities. But when it comes to the limits of the scope, the Project Management differs. The scope of a process defines what should be included in it and what is not its responsibility and ITIL as a library of good practices is defining what should be in the scope of Change Management process in a matter of only technical points of view. However Project Management is more concerned with the technical part and the management point of view – questions as who will pay and cost assessments are as equally important as the risk and impact assessment.

Politics, Main Call Flow (Activities), Roles

The organizational aspects of applying the Change Management processes concern politics, main call flows, activities and roles. These three main points are concentrated on the process of applying changes itself.

Although there are a lot of similarities in the Change Management process in the three methodologies in terms of politics they do differ significantly. ITIL is stressing more on the development related politics and PM and CMMI more on the business ones – who is paying for the changes etc. Change Management in PM authorizes, control and coordinates but does not plan, build test and implement changes itself.

The next points – activities, call flows and roles are intertwined so we are going to analyze them in parallel.

According the information presented in ***Table 2*** the basic call flow of the Change Management process is almost the same in the three methodologies. As already mentioned in the first part of the comparison similar goals define similar call flows and similar call flows define similar main activities. As a result the

following main steps of the change implementation and the according activities could be determined:

- Receiving of a change;
- Approval or rejection of the change request (CR) – activities involved: assessment of the change request, approval of the changes, assessment of the change;
- Risk and impact assessment - risk categorization;
- Scheduling the change - prioritizing, planning of the changes;
- Development of the change - assessment of the methods and rollback procedures, planning of rollback procedures;
- Application of the change - coordination of the application of the changes;
- Documenting and closing the change.

In all discussed topics ITIL offers the best coverage of different activities of the process. To implement them in the best possible way ITIL defines a wide range of roles that supports all sides of the change mechanism.

Table 2. Call flow, politics, activities and roles of the Change Management process in ITIL, CMMI and PM

	ITIL	**Project Management**	**CMMI**
Politics	- politics for not allowing unapproved changes; - synchronizing Change Management with business, project; - identifying responsibilities and responsible; - not to allow unauthorized people to apply changes; - integration with other processes in order to identify unauthorized changes and incidents due to changes; - risk assessment of the changes; - assessment of the effectiveness and efficiency. [13]	- defining who can send change request; - defining who can approve/ reject changes; - defining who can apply the change.	- defining who can send change request; - defining who can approve/reject changes; - defining who can apply the change. - defining procedures for applying and closing the change. [10]

Basic Call Flow	- receiving a change request; - CAB (Change Advisory Board) rejects or approves the change; - impact and risk assessment; - schedule of the change; - develop and test of the change; - application of the change; - documenting; - closing of the change. [1]	- receiving a change request; - CCB member is assessing the change request; - CCB member is describing the scope and impact of the change to a PM; - PM is describing the importance and impact of the change so it can be approved or rejected; - documenting the changes. [19]	- receiving a change request; - analyzing the change request; - change assessment; - assessment of the resources; - preparation of the change; - application of the change - preparation of a release; - approval of the product. [2]
Roles	- person to request the change; - sponsor of the change; - change manager - CAB; - project manager; [13]	- project manager; - sponsor; - CCB; - implementation team. [4]	- project manager; - main architect. [18]

Input/Output and metrics

The last section of the comparison refers to the one of the most important parts of the description of a process in ITIL - metrics.

The metrics are used to measure how well a process is applied and this information can be further used for improving the process if the results are not satisfactory. Taken into account these metrics, it becomes easy to understand which parts are important and which should be improved.

Table 3. Input/output and metrics of the Change Management process in ITIL, CMMI and PM.

	ITIL	**Project Management**	**CMMI**
Input/ Output	Input : change requests Output: approved/ rejected changes; changes of documentation; reports for Change Management [13]	Input : change requests Output : reports for the changes impact and decision for the change [19]	Input : change requests Output : approved/ rejected/ delayed changes

Metrics	- number of successful changes; - advantages of changes; - minimizing the downtime; - minimizing the unauthorized changes; - minimizing the emergency changes; - percentage of successful changes; - minimizing changes for which rollback procedure was applied; - minimizing failed changes; - average time for applying the changes based on emergency/ priority and type; - incidents due to changes; - percentage of accurate time estimation. [12]	- change dependencies; - time between changes; - release changes; - impact of the changes; - number of changes.	- time for the changes; - number of problems for a release; - number of meeting of CCB. [2,8]

Table 3 shows that there are metrics defined and used in the three methodologies, such as number of changes and dependencies between changes, etc. But mostly ITIL is making real importance to the metrics with which the process can be measured and improved.

As ITIL should consider the implication of performing of the change and its impact to the integrity of the whole IT infrastructure, the set of ITIL metrics is more complete than others and includes also the following metrics:

- Decreasing of the downtime of the services;
- Decreasing of the unauthorized changes;
- Decreasing the number of unplanned and emergency changes;
- Percentage of the emergency changes;
- Decreasing the number of failed changes;
- Number of successful changes.

4. Conclusions

The results of the comparison of the three methodologies presented in this paper demonstrate that the process of Change Management is quite similar in all of the three – mostly on the basic descriptions and aspects. However they do differ in other aspects as PM and CMMI are more business oriented than ITIL.

The identification of a similarities and differences between the best practices in the three discussed area and their analysis is an important contributions of this paper. Our study indicates that the organizational maturity also plays an important

role for this process. A future research could examine how this maturity should be use for better results.

The paper could help practitioners to understand the need of structured and organized process of Change Management in every IT area.

Acknowledgment. This paper is supported by Sofia University "St. Kliment Ohridski" SRF under Contract 134/2012.

References

1. Addy, R.: Effective IT Service Management. The ITIL and Beyond! Springer-Verlag 4. Berlin Heidelberg (2009)
2. Center for Business Practices: Measures of Project Management Performance and Value http://www.pmsolutions.com/uploads/pdfs/PM%20Performance%20and%20Value%20List%20of%20Measures.pdf [visited June 2011]
3. Change Control: http://en.wikipedia.org/wiki/Change_control [visited June 2011]
4. Collegiate Project Services: Project Scope & Change Management http://www.collegiateproject.com/articles/Preliminary%20Change%20Management%20Plan.pdf [visited June 2011]
5. Definition of change : http://dictionary.reference.com/browse/change [visited May 2012]
6. Grigorova K., A concept of supporting engineering Change Management, Science conference research and development of mechanical elements and systems, Srpsko Sarajevo - Jahorina, 19-20 September 2002, pp 135-140
7. Harding Roberts, M.:Project Change Management http://www.hraconsulting-ltd.co.uk/project-change-request-management-control.htm [visited June 2011]
8. InterviewQuestionAsked : interviewquestionasked.com: What are the metrics followed in project management? 2009 http://www.interviewquestionasked.com/dot-net/what-are-the-metrics-followed-in-project-management/ [visited June 2011]
9. Kaloyanova K., Some aspects of implementing ITIL, Proceedings of the 6-th Annual International Conference on Computer Science and Educaton in Computer Science, 26-29 June 2010, Fulda-Munich, Germany, pp 48-53
10. Kasse, T.: Practical Insight into CMMI, Second Edtion (2008)
11. Nikolov P., K. Kaloyanova, A Solution for Software Change Management, Proceedings of the 4th International Bulgarian-Greek Scientific Conference "Computer Science 2008", 17-20 September, Kavala, Greece, 2008, pp 478-483
12. Office of Government Commerce : Service Transition, : The Stationary Office, London, United Kingdom (2007)
13. Office of Government Commerce : The Official Introduction to the ITIL Service Lifecycle: The Stationary Office, London, United Kingdom (2007)
14. Project Management Description: http://en.wikipedia.org/wiki/Project_management [visited May 2012]
15. Project Smart. Using Change Management and Change Control Within a Project: http://www.projectsmart.co.uk/using-change-management-and-change-control-within-a-project.html (2009) [visited June 2011]
16. Qatar National Project Management: Change Control Plan Preparation Guidelines http://www.qnpm.gov.qa/english/resources/4%20Change%20Control/Change%20Control%20

Plan%20Preparation%20Guidelines.PDF [visited June 2011]
17. Role of Change Management in an Organization : www.management-hub.com: 2005-2011 http://www.management-hub.com/change-management.html [visited June 2011]
18. Software Engineering Institute: Capability Maturity Model Integration http://www.sei.cmu.edu/cmmi/ [visited June 2011]
19. Stellman, A., Greene, J. : Change Control, Applied Software Project Management (2006) http://www.stellman-greene.com/aspm/content/view/36/38/ [visited June 2011]
20. The Project Management Body of Knowledge (PMBOK): http://www.projectsmart.co.uk/pmbok.html [visited June 2012]

Modeling Z-specification in UML

Snezana Savoska,

University St. Kliment Ohridski - Bitola, Partizanska Str. BB, Bitola 7000, Macedonia
snezana.savoska@uklo.edu.mk

Abstract. Development of software models with specialized notations permits non-important details of the system to be omitted. Z-notation is a tool for specification complex systems using abstract mathematical notation. Further, Z-specifications could be converted for input on software development environments. UML is the most popular notation used today in these environments. The aim of this paper is to investigate this process of conversion of Z-specifications to UML-models. It is based on an example specification of relational model of data.

Keywords: Z-notation, UML, relational data model.

1 Introduction

Z-notation [1] has a long history. In 2002 it was accepted as ISO standard. Z-notation is based on Zermelo–Fraenkel set theory. Z-specifications could be maximally abstracts.

UML [2] has been developed by OMG as notation for object-oriented design. Object-oriented approach with UML design for software development is de-facto industrial standard used in commercially available software development environments.

Modeling in UML a Z-specification is described and discussed in this paper. As a Z-specification is used relational model of data specified in [3].

2 Relational Schema

Every Z-specification begins with some basic data types. In [3] for this purpose are used relational names, relational columns and values:

[*RNAMES*, *CNAMES*, *VALUES*]

Column names are used in this specification only for extension and they are not modeled in UML. RNAMES and VALUES are modeled initially with classes with the same names and attributes and operations not specified.

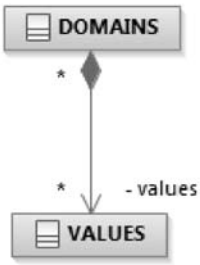
Basic types have to be modeled as classes without attributes and operations

V. Dimitrov (Editor): ISGT'2012. ISSN 1314-4855
Proceedings of the 6th International Conference on
Information Systems and Grid Technologies, Sofia, June 1-3., 2012.

sections.

$$DOMAINS == \mathbb{F}_1\ VALUES$$

The type DOMAINS is specified as power set of non-empty finite sets of values (VALUES). In UML, domains are modeled as compositions of values. The empty domain (empty composition) is included. It is not a big deviation from the specification – the model is more general.



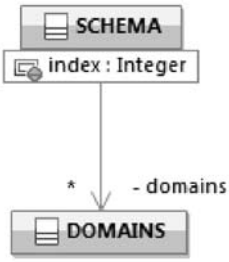
Every value could be used in more than one domain, but could not be component of any domain.

In UML, by default, only finite objects are manipulated. For example, the type Integer is in reality the subset of integers that can be represented on the computer. Here, a composition consists of finite number of elements. In UML, finite characteristic of the artifacts is not specified, because it is by default.

Relational schema is specified as non-empty sequence of domains:

$$SCHEMA == \mathrm{seq}_1\ DOMAINS$$

It is modeled as a class that has qualified association (by attribute 'index') with domains.

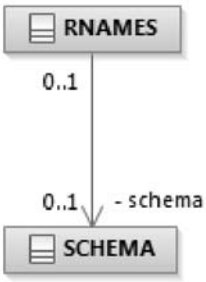


In Z-notation, sequences are indexed with naturals. A sequence in mathematics could be indexed with any countable set. So, it is not a big deviation in that case instead naturals to be used Integer for indexing. The most important property remains: domains in the schema are ordered.

$$\begin{array}{|l} \textit{DBSchema} \\ \hline db : RNAMES \rightarrowtail\!\!\!\!\!\!+\ SCHEMA \\ \hline \end{array}$$

Database schema is specified as partial function from relation names to

schemas. Database schema is specified as abstract data type. In UML, the same approach could be used, i.e. database schema could be represented as a class with attributes and operations, where an attribute (or attributes) would map relation names to schemas. This mapping in UML terms is association and it is not recommended association to be hidden with attributes. That the reason why database schema is modeled as association between relation names and schemas.



Multiplicity of this association is zero or one at both ends. The association is directed from relation names to schemas, because navigation in the other direction is not needed at this time. All associations in the diagram are minimal is sense that direction is specified only when navigation in that direction is used.

Initially, the set of relational schemas is empty:

DBSchemaInit

DBSchema

$\mathrm{dom}\ db = \varnothing$

In Z-notation above Z-schema is a constructor. In this case database schema is not modeled with a class and no constructor could be created. The association depends on classes that it connects.

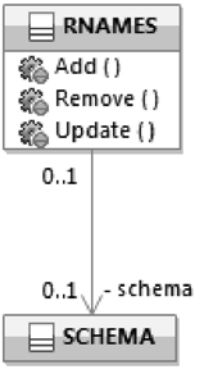
In object-oriented systems, the hypothesis is that, when the system starts, there no objects and only after the initialization new objects and links among them are created. So, by default, the set of links between relation names and schemas, initially, is empty.

$$
\begin{array}{l}
\textit{DBSAdd} \\
\hline
\Delta DBSchema \\
n?: RNAMES \\
s?: SCHEMA \\
\hline
n? \notin \mathrm{dom}\ db \wedge \\
db' = db \cup \{n? \mapsto s?\} \\
\hline
\end{array}
$$

$$
\begin{array}{l}
\textit{DBSRemove} \\
\hline
\Delta DBSchema \\
n?: RNAMES \\
\hline
n? \in \mathrm{dom}\ db \wedge \\
db' = \{n?\} \lhd db \\
\hline
\end{array}
$$

$$
\begin{array}{l}
\textit{DBSUpdate} \\
\hline
\Delta DBSchema \\
n?: RNAMES \\
s?: SCHEMA \\
\hline
n? \in \mathrm{dom}\ db \wedge \\
db' = db \oplus \{n? \mapsto s?\} \\
\hline
\end{array}
$$

Database schema has three operations: add, remove and update of relation schema. These operations manage links between RNAMES and SCHEMA. The problem now is where these operations have to be located? First location could be the association, i.e. the association between RNAMES and SCHEMA could be with class-association and operations could be placed there. In these case operations have to be static, because they create, remove and update links – instances of the association. The second possibility is to put them on RNAMES, where they would be instance operations with side effect on links of the association. This approach is better, because association, in reality, is implemented with one or two attributes in one or both participating classes (association ends are pseudo-attributes in UML) and every change of the link means change of attributes of these objects. So, this location is used in the UML-model:



There are more possibilities where to place these operations: in class SCHEMA or in another class modeling database schema. These variants go way from original concept. First of them is equivalent to the chosen one only if the association is bidirectional, but it is not true. Second variant could be implemented only with global side effects of the operations that would be hidden in the UML-model.

RNAMES is a basic type in the Z-specification, but in the UML-model it has operations. The effects of last ones are described in OCL pre- and post-conditions. For the operation Add(), pre-condition require no link to exist between the relation name and any schema: `schema->isEmpty()`. Because Add() is now instant operation, it is applied on RNAMES objects and the relation name is the first argument, by default, for Add(). Post-condition of Add() requires a link to be established between relation name and the schema (supplied as argument of Add()) objects: `schema = s`. Post-condition of operation Add() in the UML-model differs from that one in the Z-specification. In UML, post-condition refers only to changed parts and by default all other parts remain unchanged (Frame Problem of UML). In the Z-specification is clearly stated that all other links between relation names and schemas remain the same. Here, the only changed part is the link and that is why in the UML-model of Add() such a post-condition is used. In just a same way, post-conditions of the other two operations are re-mastered.

Relation names in the relational model of data are strings. In the UML-model they are objects. Every object in UML is unique, i.e. it has identity, and it follows that relation names are unique in the UML-model. When the UML-model will be further detailed an attribute of type string will be introduced in RNAMES to represent the symbolic name of the relation. This means that an invariant in RNAMES has to be introduced in the future to warranty that relation names (the string attributes values) are unique. Every RNAMES object has to have a unique name.

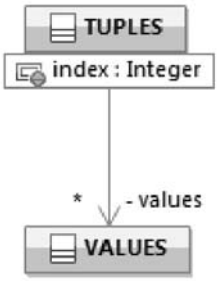
With these operations modeling of database schema is finished. In relational model of data there are logical structure and instance of the database. The first one is the database schema. The second one is a set of all relation instances. In the next section database instance is modeled.

3 Database Instance

Relational instance is a set of all its tuples in the current moment. For this purpose, tuples, first, have to be modeled. Tuple is an ordered set of values.

$$TUPLES == \mathrm{seq}_1\ VALUES$$

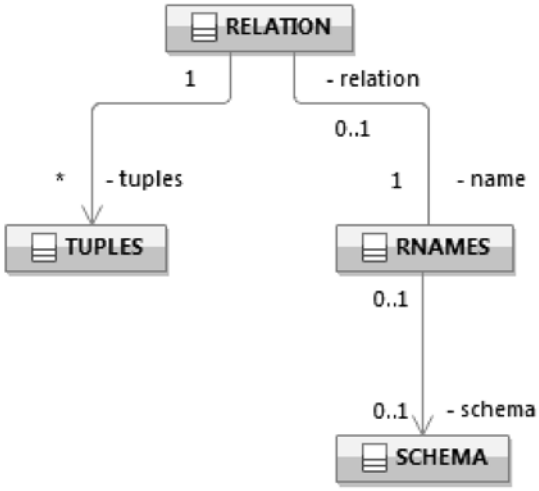
Tuples are modeled with qualified association in the same way as schema is modeled with domains. Schemas and tuples can be modeled with attributes. For example, in class TUPLES can be introduced an attribute of type VALUES with '*' multiplicity. This approach hides association between tuples and values and it is not recommended in UML.



Relation is specified with its schema and instance.

$$\begin{array}{l}
\textit{RELATION} \\
\hline
\textit{DBSchema} \\
n\colon \textit{RNAMES} \\
r\colon \mathbb{F}\ \textit{TUPLES} \\
\hline
n \in \mathrm{dom}\ db \wedge \\
(\forall t\colon \textit{TUPLES} \bullet t \in r \Rightarrow \#t = \#(db(n)) \wedge (\forall i\colon 1..\#t \bullet t(i) \in db(n)(i))) \\
\hline
\end{array}$$

Relation is modeled in the same way with the class RELATION. This class has association with a relation name. The association shows that every relation has to have relation name and transitively schema, but it is not obligatory every relation name to be bounded with a relation and vice versa. These constraints are specified with the structural notation of UML.



Every relation instance is a set of tuples. It is modeled with an association between relation and tuples. Relation instance could be empty and it is modeled with a star for multiplicity put on the association end at the class TUPLES.

Every tuple participates in exactly one relation. Relation tuples must follow relation schema. Tuples can be associated with relation (relation name) or directly with schema, but closer to the specification is to be associated with a relation (relation object) as it is modeled with two invariants:

```
tuples.values->size() = name.schema->size()
```

and

```
let n:Integer = name.schema->size()
in Set{1..n}->forAll(i | name.schema.domains[i].
values->includesAll(tuples.values[i]))
```

Tuples are lists of values from relational model point of view. In object-relational model every tuple is an object and has its own identity. This means that in object-relational model two tuples can be just same lists of values and to be different objects by their object identifiers. The Z-specification is based on the pure relational model. The UML-model accepts object-relational model: two tuples in one relation can be the same lists of values, but as TUPLES objects to be different tuples.

If pure relational interpretation is needed, in RELATION an additional invariant could be added to warrantee that all tuples in the relation are different only when they are different as lists of values. This invariant could be alternatively part of TUPLES, but that is not the way of the UML-model.

Initially, by the Z-specification, relation instance is empty:

$$
\begin{array}{l}
\textit{RELATIONInit} \\
\hline
\textit{RELATION} \\
n?: \textit{RNAMES} \\
\hline
n? \in \mathrm{dom}\ db \wedge \\
n = n? \wedge r = \varnothing \\
\hline
\end{array}
$$

This initialization of relation instance is a constructor in object-oriented terms. It can be modeled as static operation it RELATION. This operation would bind relation name with relation schema and create an empty instance for that relation. In the UML-model, constructors of all kinds are not modeled to simplify the model.

In the specification supporting Z-schema CHECK is introduced to check for tuple appliance to a relation schema. CHECK is used in relation operations: add and delete tuple. This check doubles relation invariant and is not needed. If Z-specification is extended with checks for errors this supporting Z-schema is needed to separate successful operations from unsuccessful ones, but this is not the case. In UML, there are features specially designed for errors – operation exceptions.

$$
\begin{array}{l}
\textit{CHECK} \\
\hline
db: \textit{RNAMES} \rightarrowtail\!\!\!\!\!\!+ \textit{SCHEMA} \\
n: \textit{RNAMES} \\
t: \textit{TUPLES} \\
\hline
\#t = \#(db(n)) \wedge \\
(\forall i: 1..\#t \bullet t(i) \in db(n)(i)) \\
\hline
\end{array}
$$

So, the operation Insert() simply adds and operation Delete() removes a tuple to/from the relation instance, i.e. they add/remove link between the relation and a tuple.

$$
\begin{array}{l}
\textit{RInsert} \\
\hline
\Delta RELATION \\
t?: TUPLES \\
\hline
CHECK[t? / t] \wedge t? \notin r \wedge \\
r' = r \cup \{t?\}
\end{array}
$$

$$
\begin{array}{l}
\textit{RDelete} \\
\hline
\Delta RELATION \\
t?: TUPLES \\
\hline
CHECK[t? / t] \wedge t? \in r \wedge \\
r' = r \setminus \{t?\}
\end{array}
$$

The pre-condition of Insert() requires the new tuple not to be one of the relation instance: `tuples->excludes(t)`, the pre-condition of Delete() requires the opposite: `tuples->includes(t)`.

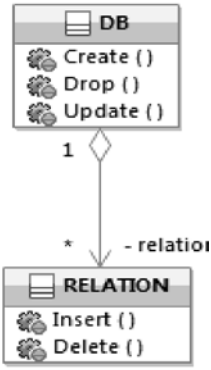
The post-conditions of Insert() and Delete() operations are:

```
tuples = tuples@pre->union(Set{t})
tuples = tuples@pre - Set {t}
```

Database instance in the Z-specification is a partial function from relation names into relations

$$
\begin{array}{l}
\textit{DBInstance} \\
\hline
instance: RNAMES \rightarrowtail RELATION \\
\hline
\forall n: RNAMES \bullet n \in \mathrm{dom}\ instance \Rightarrow n = (instance(n)).n
\end{array}
$$

Here, again the problem is to hide the association with attribute in a class or not. Following UML recommendations, the second approach is used.



Database instance is modeled with the class DB, which is an aggregate of relations. It is possible this aggregate to be empty, but every relation has to be assigned to exactly one database.

Database instance constructor is specified with:

$$
\begin{array}{|l}
\hline
\textit{DBInstanceInit} \\
\textit{DBInstance} \\
\hline
\textit{instance} = \varnothing \\
\hline
\end{array}
$$

This constructor is not modeled following above mentioned reasons.

$$
\begin{array}{|l}
\hline
\textit{DBAdd} \\
\Delta \textit{DBInstance} \\
n? : \textit{RNAMES} \\
r? : \textit{RELATION} \\
\hline
n? \notin \operatorname{dom} \textit{instance} \wedge \\
n? = r?.n \wedge \\
\textit{instance}' = \textit{instance} \cup \{n? \mapsto r?\} \\
\hline
\end{array}
$$

$$
\begin{array}{|l}
\hline
\textit{DBRemove} \\
\Delta \textit{DBInstance} \\
n? : \textit{RNAMES} \\
\hline
n? \in \operatorname{dom} \textit{instance} \wedge \\
\textit{instance}' = \{n?\} \lhd \textit{instance} \\
\hline
\end{array}
$$

There are supporting operations DBAdd and DBRemove in Z-specification. They add/remove relation instance to/from database instance. The real operations are DBSCreate and DBSDrop.

DBSCreate

$\Delta DBInstance$

$n?: RNAMES$

$n? \notin \mathbf{dom}\ instance \wedge$

$(\exists r: RELATION \bullet n? = r.n \wedge r.r = \varnothing \wedge$
$instance' = instance \cup \{n? \mapsto r\})$

DBSDrop

DBSRemove

DBRemove

DBSCreate binds a relation name with a schema and create an empty instance for the newly created relation. DBSDrop removes relation schema and its instance. In the UML-model, relational schema and its instance are associated through the class RELATION. This approach is used the relational model – there is no clear notation for relation schema and relation instance as in object-oriented approach for class and class extent. Supporting Z-schemas are not modeled – they are included in modeling of DBSCreate and DBSDrop in the class DB. The last ones are modeled with the operations Create() and Drop(). In UML, it is possible to simplify the operations names, because they are local in the class. In Z-notation, Z-schema names are global and have to be unique. For that reason in Z-specifications a naming convention for Z-schemas has to be used.

Create() pre-condition is: `relation.name->excludes(n)` and its post-condition is: `relation.name->includes(n) and n.relation.tuples->isEmpty()`.

Drop() pre-condition is: `relation.name->includes(n)` and its post-condition is: `relation.name->excludes(n)`.

There is one more operation on database instance – DBUpdate.

$COLUMNS == \mathbf{seq}_1\ CNAMES$

NSCHEMA

$s: SCHEMA$

$ns: COLUMNS$

$\#ns = \#s$

DBUpdate do not use supporting Z-schemas and it is directly modeled as operation Update() in class DB. Update() pre-condition is: `relation.name->includes(n)` and its post-condition is: `relation->includes(r) and r.name =`n.

Finally, in the Z-specification, there are two more Z-schemas for its extension with named relation columns.

$$
\begin{array}{|l}
\hline
\textit{DBUpdate} \\
\Delta DBInstance \\
n?: RNAMES \\
r?: RELATION \\
\hline
n? \in \mathrm{dom}\ instance \wedge \\
n? = r?.n \wedge \\
instance' = instance \oplus \{n? \mapsto r?\} \\
\hline
\end{array}
$$

This extension is not included in the UML-model, because the model has to re-mastered and step by step modeling would be lost.

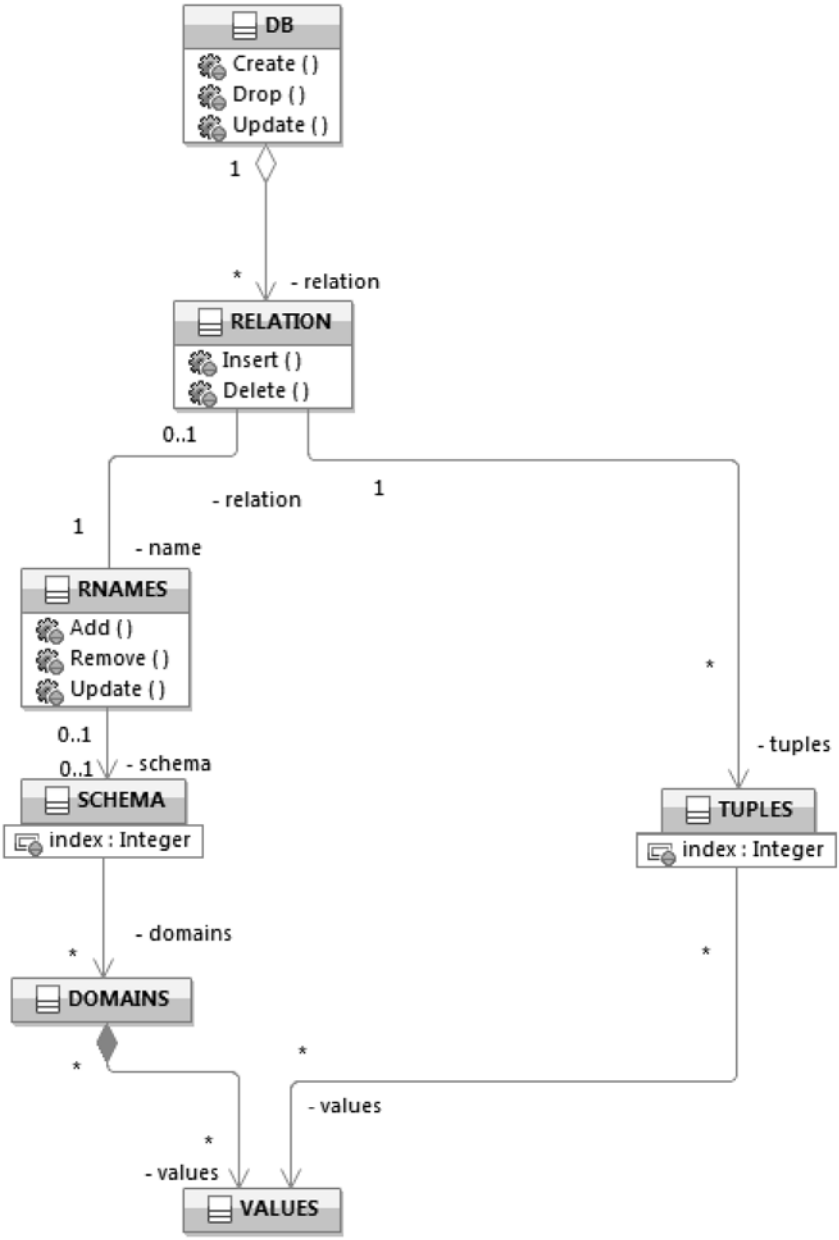
4 Conclusion

The most frequently used approach in UML-modeling of Z-specifications is to unhide hidden associations. Z-notation is a tool for specification of abstract data types. In Z-specification associations are clearly defined with functions. In UML, association may be hidden with attributes and its concept becomes hidden for the reader that is why it is recommended associations to be used instead of attributes.

The UML-model presented here is very abstract – it needs of further re-mastering. For example, the classes RNAMES and SCHEMA form relational database catalog. The last one could be implemented with relations, i.e. the catalog has to be described in terms of relations with self-describing initialization.

The Z-specification is based on the concept for relational model of data described in [4]. There semantics of the model is based on the domains. In the implementations of relational model like DB2, Oracle and so on, this concept is not well supported. These implementations are based on SQL that has been developed as a common query language for relational and hierarchical databases, and as result of that semantics of relational model has been lost.

The whole UML-model as class diagram is:



Relations with column names are called in [4] ‘relationships’. They are specified in [5] and Z-specification there can be used for development of UML model at higher level of abstraction.

Finally, relational model is packed with query language; relational algebra is proposed in [4] as such a language. Z-specification of relational algebra is given

in [6]. The last one is the natural direction for further modeling of relational model in UML.

References

1. ISO/IEC 13568: 2002 (E) Information Technology. Z Formal Specification Notation. Syntax, Type System and Semantics, www.iso.org.
2. Unified Modeling Language, OMG, http://www.uml.org
3. Dimitrov, V.: Formal Specification of Relational Model of Data in Z-Notation, Proc. of XXXIX Conf. of UMB, 178-183 (2010)
4. Codd, E.F.: A Relational Model of Data for Large Shared Data Banks. CACM vol. 18, no 6, 377--387 (1970)
5. Dimitrov, V.: "Relationship" Specified in Z-Notation, Physics of Elementary Particles and Atomic Nuclei, Letters, Vol. 8, No. 4(167), 655—663 (2011)
6. Dimitrov, V.: Formal Specification of Relational Model with Relational Algebra Operations, Proc. of Fifth Int. Conf. ISGT, 25-28 May 2011, Sofia, pp. 57-74.

Logical Design for Configuration Management Based on ITIL

Hristo Kyurkchiev[1], Kalinka Kaloyanova[1]

[1]Faculty of Mathematics and Informatics, Sofia University, 5 James Bourchier blvd., 1164, Sofia, Bulgaria

hkyurkchiev@fmi.uni-sofia.bg, kkaloyanova@fmi.uni-sofia.bg

Abstract. The focus of this paper is the Configuration Management Database (CMDB) as seen by Information Technology Infrastructure Library (ITIL). It aims at providing a conceptual and logical model for this database. The core theoretical CMDB concepts are considered and also many practical problems are discovered. Having uncovered the main obstacles and problems that an implementer faces, valid solutions are provided. As a result a logical model of a CMDB is proposed. Its practical value for implementation has been tested with real life data and satisfactory results are reached.

Keywords: IT Services, ITIL, Configuration Management Database, Knowledge Management, Logical Database Model

Introduction

IT services are one of the sectors in the world's economy that practically grows by the minute. In this environment, a knowledge reference for managing IT services was created by the Office of Government Commerce (OGC) [10]. Ever since its first version the Information Technology Infrastructure Library or ITIL has been pushing the envelope to making better, more reliable IT services through the adoption of some basic guidelines and best practices. As of the third and last version of the library this is done through separating the service lifecycle into phases and the phases themselves into processes [10]. Configuration Management is one of the most important processes in ITIL. Through its interconnections with other processes e.g. Change Management, Incident Management, Problem Management etc., it plays considerable role in providing a sound and stable IT service. The Configuration Management Database (CMDB) is in the core of this process and the acronym CMDB is even better known than ITIL [12]. The broad scope of the definitions of the concepts connected with this database, coupled with the fact that there already is an abundance of commercial-grade solutions

which are making an effort to implement it in different ways, only introduce uncertainty and confusion for the people who are trying to introduce a CMDB into their organizations.

ITIL's Versions and the CMDB

ITIL v1 and v2 are focused on the processes that facilitate IT Service Management. ITIL v1 contains thirty books, whereas ITIL v2 only has eight (Service Support, Service Delivery, ICT Infrastructure Management, Security Management, The Business Perspective, Application Management, Software Asset Management, Planning to Implement Service Management) [1], [4]. However, since ITIL v2 is a consolidation of ITIL v1 both versions are largely the same [4]. The Configuration Management process is located in the Service Support section [1]. Version 3 on the other hand focuses on the service lifecycle. It identifies five lifecycle phases – Service Strategy, Service Design, Service Transition, Service Operation and Continual Service Improvement [10]. The Service Asset and Configuration Management process is in the Service Transition Phase [10]. However, no matter which version of ITIL is implemented, the Configuration Management (CM) process and more precisely the CMDB is in the core of the implementation and a deep understanding of all concepts that concern this database is necessary [4].

How has CMDB evolved in order to handle the dynamic IT environment? A comparison between the different versions of the configuration management process can be seen on Table 1.

Table 1. Configuration Management concepts in the different versions of ITIL

	ITIL v1 and v2 [1]	**ITIL v3 [10]**
Focus	Processes	IT Service Lifecycle
CM Process	Configuration Management	Service Asset and Configuration Management
Responsibilities	CM provides a logical model of the infrastructure or a service by identifying, controlling, maintaining and verifying the versions of Configuration items (CIs) in existence.	CM ensures that selected components of a complete service, system or product (the configuration) are identified, baselined and maintained and that changes to them are controlled.

Configuration Item (CI)	Component of an infrastructure – or an item, such as a Request for Change, associated with an infrastructure – that is (or is to be) under the control of CM. Configuration Items (CIs) may vary widely in complexity, size and type, from an entire system (including all hardware, software and documentation) to a single module or a minor hardware component.	Any Component that needs to be managed in order to deliver an IT Service. Information about each CI is recorded in a Configuration Record within the Configuration Management System and is maintained throughout its Lifecycle by Configuration Management. CIs are under the control of Change Management. CIs typically include IT Services, hardware, software, buildings, people, and formal documentation such as Process documentation and SLAs.
Configuration Management Database (CMDB)	A database that contains all relevant details of each CI and details of the important relationships between CIs.	A database used to store Configuration Records throughout their Lifecycle. The CM System maintains one or more CMDBs, and each CMDB stores Attributes of CIs, and Relationships with other CIs.
Configuration Management System (CMS)	A software product providing automatic support for Change, Configuration or version control (Configuration Management Tool is used instead of Configuration Management System).	A set of tools and databases that are used to manage an IT Service Provider's Configuration data. The CMS also includes information about Incidents, Problems, Known Errors, Changes and Releases; and may contain data about employees, Suppliers, locations, Business Units, Customers and Users. The CMS includes tools for managing data about all CIs and their Relationships. The CMS is maintained by CM and is used by all IT Service Management Processes.

Despite the fact that in the third iteration the terms get more detailed there are no considerable differences between the three versions of the IT infrastructure library. Several main conclusions can be drawn from this theoretical survey:

- Configuration Management is not simply a process for recording assets' locations and their accountability – the configuration infrastructure includes items that are outside of the scope of the purely technical aspect of the service and by doing so it poses a considerable obstacle to the person willing to implement a real life software solution. The need for this software system to be able not only to provide accounting information about the assets, but also to acquire new knowledge about the managed service, its configuration,

sustainability, and abilities for future development make it an even harder task. The vital importance of this knowledge extraction through the CMDB is stressed by the introduction of the new Knowledge Management process, introduced in ITIL v3. The low maturity of this process and its vague definition, which sometimes copies the one for Configuration Management, suggests that a sound CM process and CMDB are the keys for the future knowledge discovery and extraction in the ITIL framework.

- The implementation of a successful Configuration Management process starts with the development of a sound CMDB solution – since the database is the core of the CM process its stability is correlated to the stability and robustness of the whole process. The standard process for database development is used to create a sound CMDB solution. It starts with the building of a conceptual and logical model. To enable this the CIs, their parameters and types should be carefully researched, as well as the ways in which they can be interconnected.
- The maturity of the Configuration Management process is dependent on the maturity of the other processes – it is not possible for the CM process to exist in a vacuum and without support from the other processes it cannot reach its full potential. Its closest connections are with the Incident Management, Problem Management, Change Management, Release and Deployment Management, Availability Management, and Finance Management processes. For example, a sound Configuration Management solution would provide Change Management with enough information about what an impact a change would have on the infrastructure. Also, in order for the CI records in the CMDB to be kept up to date a sound Change Management process is required. This holds true for all of the aforementioned symbiotic relationships.
- Although of significant importance Configuration Management is not the universal remedy – judging by the business goals that ITIL v3 outlines for this process [10] one can reach a conclusion that CM is all he or she needs. This, however, is not the case, as can be easily seen by its interconnectedness.

CMDB in Practice

We have already established that the CMDB is not just a database. However, prior to the introduction of ITIL v3 the idea that the practitioners had, which probably stemmed from ITIL v2 [1] was that it was a single database, which in the eyes of the specialists in the field means a single information source, usually of relational origin. An example of such a vision is shown by Spafford [16] who states that the CMDB is a relational database, which serves as a "nerve" center for the management of IT services.

Several authors also agree with the aforementioned statement that the CMDB

is not just a database [9], [12], [13]. This is in line with the vision of ITIL v3, according to which the CMDB is an integrated source of data, which acquires information from many other databases. In order to avoid confusion O'Donnell [13] and also Marquis [9] prefer to use a different term for this vision of the CMDB - Configuration Management System in one case and Configuration Management Databank in the other. In this manner, if the CMDB is considered as one of the numerous databases from which the integrated CMDB acquires its information, then the view for a single relational database, as was the previous idea for the CMDB is completely valid. The same holds true if we think for the integrated CMDB in the terms of Configuration Management System. Marquis [10] even goes as far as to suggest that the CMDB is not just a databank, but also a metabase (metadatabase, metadata repository). He insists that the key to the integrated CMDB is the joining of the metadata for and from all data sources into one database, which in his eyes is the CMDB. If we imagine that there are numerous available data stores, which define the configuration of the IT service, and which can be of various types and granularity, then the CMDB would contain information about all these data stores and would provide the user with an integrated, run-time generated information from them.

These ideas about the CMDB lead to different visions about its implementation too. On one hand, there are those, who like Spafford [16] suggest using the relational model for the CMDB, as it has become, in recent times, a synonym of databases. He supports this suggestion by offering some classification of the most important attributes that a CI has. On the other hand, if we consider the CMDB as an integrated source from disparate sources, it becomes clear that the relational model is not a suitable option for implementation. This is further stressed by the data that is to be saved and extracted – usually connected with different business metrics, such as load, availability, efficiency, effectiveness, etc. All these suggest the usage of the dimensional model, which as Marquis [10] points out is suitable for on line analytical data processing as opposed to the on line transactional data processing typical of the relational model. Both views have been used for real life CMDBs: OpenCMDB is a relational database, and RapidCMDB is an example of a CMDB using the dimensional data model. Which approach to choose depends largely on the organization size and needs, as well as on its capacity and financial conditions.

Conceptual and Logical Model of a CMDB

So far, we have established that the CMDB should save information about the CIs and their interconnections. As simple as it may sound this poses two problems for the implementation:

1. The CIs are of great variety of types and unite different kinds of entities – hardware and software components, people, buildings, documents, etc.

2. The relationships between the CIs are not clearly defined in ITIL and in practice between two arbitrary CIs an arbitrary type of relationship can exist. This poses two additional obstacles:
   a. What kind of relationships should the system support?
   b. How to control the relationships so that only the possible and logical relationships can exist between two specific configuration items?

We tried to find the solutions of these problems in the process of building a conceptual and logical model for the CMDB.

The simplest conceptual model of a CMDB follows directly from the definition of this database [10] and can be seen on Figure 1. It was suggested by Schaaf and Gögetap [15] and although it is of little practical value due to its high conceptual standing it is useful in that it shows visually the definition of the CMDB. Every CI and relationship between CIs has a type associated with it, and every instance of a relationship binds two CIs in a relationship. The beauty of the model is in its simplicity. It is perhaps the only true conceptual model, as every other model adds to it. Lead by this we would look at every other model as a logical one, as it would further define the concepts and by doing so would be of a more concrete and less abstract conceptual level. In order to reach such a model we would address the two issues of the types of CIs and the relationships between them.

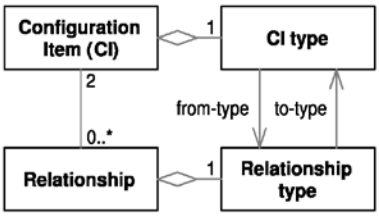


Fig. 1. Conceptual CMDB Model

Every basic classification of CIs identifies physical and logical ones [2], [3]. However, this is too simple and of little practical value. Betz [2] gives another classification, which separates the CI into three classes – CIs, Operational CIs, and Production CIs. These are abstract classes and no object from them can exist. The CI class is the broadest one and every other is inherent from it. The Operational CIs are the CIs involved in day-to-day business processes; they can be measured and are a primary entity in the Service Management workflow [2]. The Production CIs are the ones that are responsible for the actual delivery of the service. As a general rule, in order for a CI record to be in the CMDB it should be important enough to fall under formal Change Management. This classification is more defined but lacks some details that are essential to the modern view of the CMDB. However, it provides for a good basis and we used it for the development of our own classification, which is given on Figure 2.

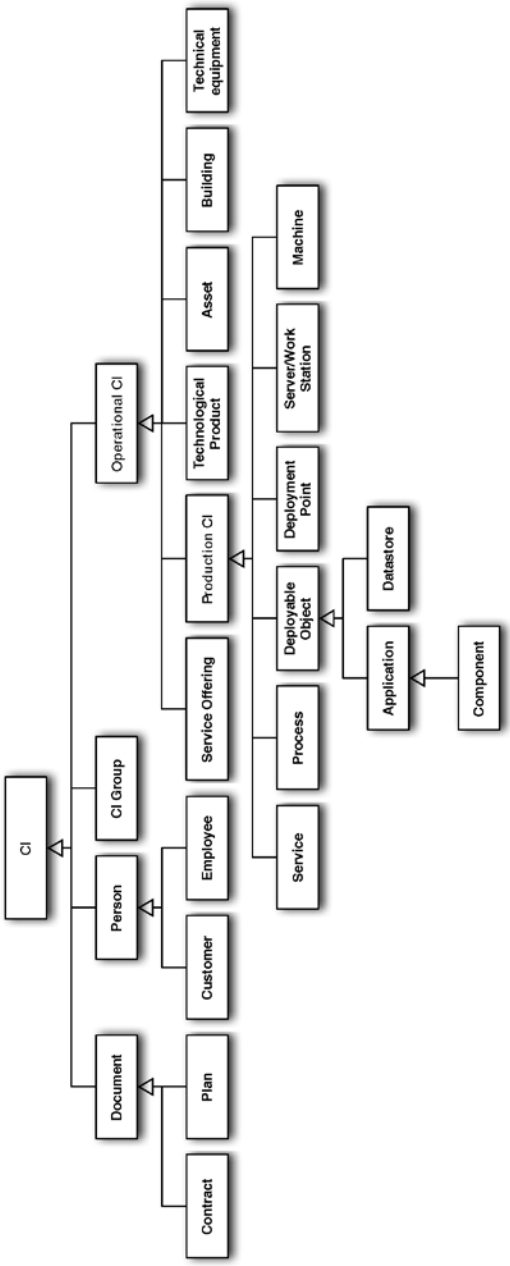


Fig. 2. A Hierarchy of CIs

Several important additions to the model that Betz [2] suggested are made:

- The plan type of document, which is important as it can often include

measurements that can be used as comparison points for different aspects of the service;

- The person type of CI, since according to OGC [14], and McAlpin [11] people involved in the service operation should be treated as a CI. This stems from the fact that personnel is one of the main expenses that an IT organization has and that many problems and incidents may be the result of human error or incompetence. We included two different roles, which a person can have – the customer role and the employee role. These are not necessarily disparate sets, as an employee, regardless of his position in the organization may act as a customer for a system that is hosted by a department different from the one he or she is appointed in.
- The building type of operational CI, which is important as with the joining of the Asset Management process and the Configuration Management process the buildings should be accounted for. The necessity of a separate CI type from the one for asset is suggested by the fact that buildings can be leased and leased objects are not considered assets. Moreover, buildings and any type of space can encompass different physical CIs and as a result need to be managed and controlled, for example, for recovery and contingency measures.
- The technical equipment type of operational CI, which include server racks, disk warehouses, etc. Again, these could be leased and do not count toward traditional assets.

Having reached a hierarchy of the CIs the next remaining question is the possible relationships between the CIs. The main relationships archetypes are dependency, ownership, and enabling. These archetypes can be further decomposed into concrete relationship types as has been done by BMC [3] and Betz [2].

In his model Betz [2] included only two relationship types – ownership and dependence. He also limited the possible relationships between the CIs by allowing only CIs of the same type to exist. This greatly simplifies the model and the implementation. However, we consider that a logical model should be as close to reality as possible so in our model we allow cross-type CI relationships. As a result, the decision about which relationships to be used in a certain system remains a part of its implementation.

Having answered the two main questions we built a logical model for a CMDB with the use of E/R notation, where the entities are represented by rectangles and the relationships' multiplicity is represented by arrows with a straight arrow end meaning one and a trident arrow end meaning many. In order not to crowd the model we have included only the three main archetypes of relationships – dependence, enabling, and ownership. Also, besides the items from the CI hierarchy included are some entities that have not been discussed so far, the examples of which include Program, Project, Release, Problem, (Request for) Change, Incident, Event, Service Request, Risk, and Known Error. Although these entities are primarily the concern of other ITIL processes, without data

and information about them the CMDB would not be complete. Their addition strengthens the integrated nature of the database and makes it more capable for knowledge discovery and mining. The result may be seen on Figure 3.

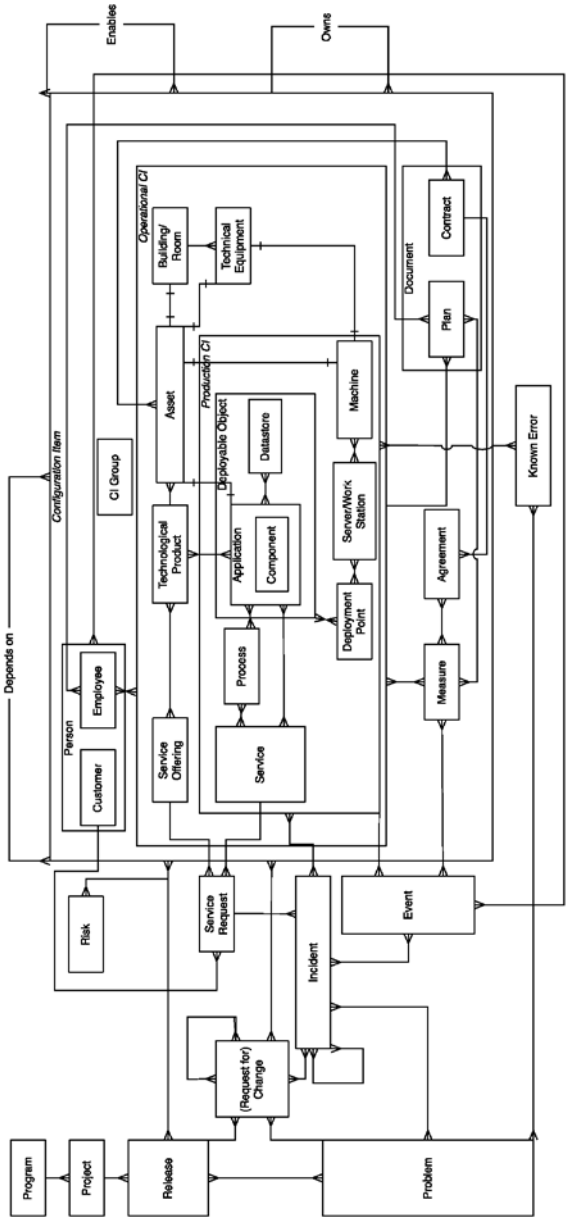


Fig. 3. Logical Model of a CMDB

In order to validate the model, we have developed a test physical model of a CMDB. In it we concentrated on the needs of the small to medium size companies, since there are plenty of commercial grade solutions for the big enterprises (such as BMC's Remedy, IBM's Tivoli Change and Configuration Management Database, Microsoft's System Center and so on). Moreover, the purpose of the creation of the model was to ensure that the concepts we have put into our logical model provide for a sound CMDB solution. The physical model itself implemented only a subset of the logical one, since the needs of the small to medium size businesses are met even without the full scope of the model. We wrote the SQL scripts for the actual creation of the physical model and filled it up with example data, based on real life IT infrastructure details. Then we ran some tests on the database based on real life scenarios. The database proved itself able to handle the necessities of an IT service provider's day-to-day needs, as far as Configuration Management concerns go. It provided some simpler reporting compared to the mentioned commercial solutions, but it was also able to do it at lower production and performance costs and in a more user-friendly way. This also held true for the infrastructure related queries the likes of available storage, infrastructure load, occurring of incidents, etc. These results lead us to believe that the model can be the basis for a full-fledged solution.

Conclusions and Future Work

The building of a logical model is the first step towards the implementation of a sound database solution. The real implementation of ITIL principles and particularly CMDB depends on many details [7] but constructing a model, which is further followed, we confirm once more the significance of the theoretical foundation for improving the modeling process [8], [6]. The model we are suggesting is based on the definitions of ITIL and reflects ideas of various ITIL practitioners. Thus we consider it to be a solid ground on which to elaborate at a later time. However, improvements are always possible.

Additions to the model are possible especially with next revisions of the ITIL. Other sources of such additions are the demands of the real business processes [5]. The new Knowledge Management process is also an interesting source of new requirements for the CMDB since it so heavily relies on it. And knowledge extraction poses quite a few challenges to modern databases. Future work also includes refining the physical model and building working CMDB management software around it.

Acknowledgment. This paper is supported by Sofia University "St. Kliment Ohridski" SRF under Contract 134/2012.

References

[1] Berkhout, M. et al., ITIL: Service Support. [CD edition] London, UK: Crown, 2003.

[2] Betz, Ch. (2005, August 13) A Data Architecture for IT Service Management [On-line]. Available: http://erp4it.typepad.com/erp4it/files/a_data_architecture_for_it_service_management.pdf [May 22, 2011].

[3] BMC Software., Step-by-Step Guide to Building a CMDB, BMC Software, 2008.

[4] Dubey, S. (2008, December 3) ITIL v2 vs. ITIL v3: A Newbie's Guide [On-line]. Available: http://www.saurabhdubey.org/itilv2-v3.pdf [April 1, 2011].

[5] Grigorova K., Process manipulation within DBMS environment, CompSysTech'2002, Sofia, pp.III.28-1-6

[6] Grigorova K. Visualization of database process models, Известия на Съюза на учените - Русе, Серия 5 "Математика, информатика и физика", т. 5, 2005, 67-72

[7] Kaloyanova K., Some aspects of implementing ITIL, Proceedings of the 6-th Annual International Conference on Computer Science and Educaton in Computer Science, 26-29 June 2010, Fulda-Munich, Germany, pp 48-53

[8] Maneva N., Kr. Kraychev, Kr. Manev. Mathematical Models-Based Software Modernization. Mathematika Balkanika J., New Series, vol. 25, No 1-2(2011), pp.131-144

[9] Marquis, H. (2010, April 16) Configuration Management for the Rest of Us [On-line]. Available: http://www.itsmsolutions.com/newsletters/DITYvol6iss15.htm [May 9, 2011].

[10] Marquis, H. (2010, April 30) Enterprise CMDB [On-line]. Available: http://www.itsmsolutions.com/newsletters/DITYvol6iss17.htm> [May 10, 2011].

[11] McAlpin, T. (2007). Can People Be Configuration Items? [On-line]. Available: http://rjrinnovations.com/rjr_docs/Viewpoint Article -Can People Be Configuration Items.pdf [May 27, 2011].

[12] Messineo, D., White Paper: Myths of the CMDB. CA Services. 2009.

[13] O'Donnell, G. (2010, May 10) CMS Pockets of the Truth or CMDB Unified Ambiguity? [On-line]. Available: http://blogs.computerworlduk.com/infrastructure-and-operations/2010/05/cms-pockets-of-the-truth-or-cmdb-unified-ambiguity/index.htm [May 9, 2011].

[14] Office of Government Commerce, ITIL v3: Service Transition, Norwich, UK: The Stationary Office, 2007.

[15] Schaaf, T., Gögetap, B. Requirements and Recommendations for the Realization of a Configuration Management Database [On-line]. Available: http://wwwmnmteam.informatik.uni-muenchen.de/pub/Publikationen/gosc07/PDF-Version/gosc07.pdf [May 20, 2011].

[16] Spafford, G. (2006, April 1) Demystifying the CMDB [On-line]. Available: http://www.spaffordconsulting.com/Demystifying the CMDB.html [May 9, 2011].

An Overview of the Moodle Platform

Vladimir Dimitrov

Faculty of Mathematics and Informatics, Sofia University, Sofia, Bulgaria

cht@fmi.uni-sofia.bg

Abstract. This paper presents an overview of the Moodle platform and its installation at moodle.openfmi.net, which serves the needs of the Faculty of Mathematics and Informatics of Sofia University.

Keywords: Faculty of Mathematics and Informatics, Moodle

1 Introduction

Moodle [1] stands for Modular Object-Oriented Dynamic Learning Environment. This is a free, open source e-learning platform. According to [2], there are more than 66,000 Moodle installations all over the world, serving more than 6,000,000 courses to more than 58,000,000 users. The system is also quite popular in Bulgaria and has been installed in many universities and schools [3].

Moodle was created by Martin Dougiamas and the first version was released on 20 August 2002. Here is what Martin himself says about the name of the system:

"When coming up with a name for my system I remember sitting down one evening trying to come up with a word that: (1) was an acronym (because I like hidden meanings) (2) was a word you could say easily (3) was not common on the internet (so searches could find it) (4) had a domain name free. I played around with words and whois for a few hours before finally deciding on Moodle and registered moodle.com. The fact that "moodle" actually had a meaning of its own which made sense was the main reason why it won over other combinations. The system has never had another name, although originally the M in Moodle was "Martin's" not "Modular". [4]

Currently, most of the development of Moodle is done by Moodle Pty, Ltd [5]., a company located in Perth, Australia. It also relies heavily on the Moodle Community (more than 1 million people, who interact and provide ideas, source code, information and free support) and the Moodle Partners [6], who provide most of the funding.

V. Dimitrov (Editor): ISGT'2012. ISSN 1314-4855
Proceedings of the 6th International Conference on
Information Systems and Grid Technologies, Sofia, June 1-3., 2012.

2 Moodle

As an e-learning platform, Moodle provides functionality in several areas, with some of the most significant being:

- Course management
- Assignments management – includes submission, grading, feedback
- On-line tests – includes both performing and developing tests
- Grading management
- Calendar management
- File management
- Messaging support
- Forums and chats
- Wiki support

A very important aspect of Moodle is that it is a free, open-source system and every organization can adapt the system to better serve its needs. Also, the platform comes with options for integration with already existing systems (for example, it supports LDAP and Jabber). It also provides PayPal support.

Moodle is developed in PHP and typically runs on a MySQL database, although it also supports PostgreSQL, MSSQL, Oracle and SQLite [7]. The primary supported web servers are Apache and Microsoft IIS, although Moodle's documentation says that the system should also work on lighttpd, nginx, cherokee, zeus and LiteSpeed [7]. It also states that a realistic minimum for its hardware requirements is 5 GB of hard disk space and 1 GB of RAM for each 10-20 concurrent users.

Many hosting providers provide Moodle support and one can install the platform without having knowledge of web server administration. There are also dedicated Moodle hosting providers, which specialize in the platform and may be a good choice for non-specialists, who are mainly interested in Moodle.

As mentioned above, developers can extend Moodle. The best way to do this is by creating plugins, which introduce new functionalities. The system has a highly modular design and many types of plugins are supported.

Developers can extend Moodle's modular construction by creating plugins for specific new functionalities. Moodle supports many types of plug-ins, for example:

- Activity modules – Introduce new activities to Moodle.
- Authentication Plugins – Can be used to define custom methods of authentication.
- Blocks – allow the creation of multi-purpose blocks that contain arbitrary HTML code.

- Filters – transform content before it is delivered to the user.
- Repository Plugins – allow Moodle to access content stored in external repositories.
- and many others.

There is also a rich database of plugins developed by third-parties and made available through the Moodle website [9]. If your organization needs some specific functionality, there is a good chance that someone has already written a plugin for it.

Moodle has a large community, which provides information and support. Typically one can get answers to their questions relatively quickly on the Moodle forums [10]. There are also rich options for import and export of information from/to other systems. If you have an already existing system and want to migrate to Moodle, it is strongly recommended that you consult Moodle's documentation to see if the data from your system can be easily imported.

Moodle provides certification through the Moodle Partners [11]. Currently there is only one certificate that can be obtained – the "Moodle Course Creator Certificate". As stated on the official site: "The MCCC is a way for you to demonstrate your skills in using Moodle as a teacher. … The content of the certification is designed by the Moodle community at moodle.org where you can find discussions, documentation and preparation guides. The certification process is administered by Moodle Partners and costs between AU$200-AU$800 in your local currency (Moodle Partners in each country set the local rate)" [12].

3 Moodle.openfmi.net

Moodle was first installed at our faculty by Kalin Georgiev and Trifon Trifonov in 2004. Since then the system has served more than 4500 users (more than 2000 of them are active) and stores more than 400 courses. Currently we are running Moodle 2.1.x

Through the years, our teaching staff has gained a lot of experience with the platform. We have identified several areas in which the system may be improved or expanded, as to better serve our own needs. We aim to:

1. Better align the system to our own educational processes.
2. Always search for and implement new ways to increase effectiveness and efficiency in our everyday activities.
3. Make sure that we apply the confirmed best-practices and provide our students with the best possible learning conditions.

Research on such improvements was carried by a team of members of the Computer Informatics department, under the "Automated Tools for Assisting the Educational Process in Informatics" project, funded by the Scientific Research Fund of Sofia University [13].

Acknowledgment. This paper is supported by Sofia University "St. Kliment Ohridski" SRF under Contract 134/2012.

4 References

1. Moodle Platform (Official Web Site) http://moodle.org/
2. Moodle Statistics. http://moodle.org/stats (last visited 16.04.2012)
3. Moodle's Bulgarian Section. http://moodle.org/course/view.php?id=43
4. Dougiamas, M. "Moodle origins". http://moodle.org/mod/forum/discuss.php?d=27533&parent =129848 (last accessed 16.04.2012)
5. Moodle Pty, Ltd. http://www.insideview.com/directory/moodle-pty-ltd
6. Moodle Partners. http://moodle.com/partners/
7. Installing Moodle. http://docs.moodle.org/22/en/Installing_Moodle (last visited 16.04. 2012)
8. Moodle Developer Documentation: Make a New Plugin. http://docs.moodle.org/22/ en/ Developer_documentation#Make_a_new_plugin (last visited 16.04.2012)
9. Moodle Plugins Directory. http://moodle.org/plugins/
10. Moodle Forums. http://moodle.org/forums/
11. Moodle Certification. http://moodle.com/certification/
12. Moodle Certificates. http://certificates.moodle.com/
13. Semerdzhiev, A., T. Trifonov, M. Nisheva. "Automated Tools for Assisting the Educational Process in Informatics". Published in the Proceedings of the International Conference on Application of Information and Communication Technology in Economy and Education (ICAICTEE) 2011, Sofia, Bulgaria, (pp. 429–434). ISBN 9789549224733 (in Bulgarian)

Planned Improvements for `moodle.openfmi.net`

Atanas Semerdzhiev[1]

[1] Sofia University, Faculty of Mathematics and Informatics, Sofia, Bulgaria

asemerdzhiev@fmi.uni-sofia.bg

Abstract. moodle.openfmi.net is a Moodle installation at the Faculty of Mathematics and Informatics of Sofia University. Through the years, we have gained significant experience with the system and have identified several areas, which can be improved or extended in order to better support the specific needs of our educational process. This article describes the improvements, which we plan to introduce to the system in 2012.

Keywords: FMI, Moodle

1 Introduction

Moodle was first installed at The Faculty of Mathematics and Informatics (FMI) of Sofia University (SU), during the summer of 2004 by two of our colleagues – Kalin Georgiev and Trifon Trifonov. Soon the full potential of the platform was recognized and many courses were transferred to the system. It also obtained its current URL [2]. Today there are several Moodle installations at FMI and each serves a different purpose. Our system (moodle.openfmi.net) has more than 2100 active users and our database stores more than 400 courses.

Through the years, the teaching staff at SU-FMI, have gained significant experience in applying Moodle to different aspects of the educational process. We have identified several areas in which the system can be improved or extended in order to better support our specific needs. This article describes the improvements and new functionalities, which we plan to introduce in the system until the end of the year.

2 Planned Improvements

2.1 Automatic Homework Testing and Plagiarism Detection

In 2011, a team of colleagues from the Computer Informatics department of SU-FMI started working on the "Automated Tools for Assisting the Educational Process in Informatics" project, funded by the Scientific Research Fund of Sofia University. Five work packages were defined:

1. Analysis of the capabilities of the existing moodle.openfmi.net platform.
2. Integration of existing systems to the e-learning platform.
3. Research and development of methods and algorithms for automatic processing of the submitted solutions to programming assignments.
4. Development and integration of new and existing components and systems to aid the educational process.
5. Dissemination and popularization of the results of the project.

In [3] we gave a description of our work on work package 3, which was focused on the development of tools for automated testing of submitted programming assignments and plagiarism detection. In the winter of 2011, we tested an initial prototype of these tools on assignment submissions written in the C++ and Scheme programming languages. A more detailed account of the results will be presented in another article. Here I will note two important points:

1. The obtained results made it clear that these tools can help us quickly separate working from non-working solutions and focus on the problematic ones. Therefore, a lot of time could be saved and put to better use. As we wrote in [3], a human teacher always reviews the submitted code, but the time savings are nonetheless significant. Also, they enable us to run more sophisticated tests on the submitted solutions. On the other hand, as expected, the tools offer less help in the task of locating the problem and writing a short feedback for the student.
2. The plagiarism detection tools found several cheating attempts, including some that we would have otherwise missed. Proving that a plagiarism attempt was made (which is necessary before taking any action) turned out to be a difficult task which required additional tools and information.

One of the tasks in the future development of the automatic testing and plagiarism detection system is to integrate it with Moodle. The goal is not only to provide a better user interface (currently it is not suitable for end-users), but also to seamlessly integrate it in the same environment, which we use for our day-to-day teaching activities.

2.2 LDAP Synchronization

Currently moodle.openfmi.net uses LDAP to retrieve information about its users and for authentication purposes. Moodle comes with a built-in LDAP support, but we have found out that we need additional functionality.

What is more specific in our situation is that (1) the information in the directory server is not always consistent (for example we have some older entries, which are organized differently) and (2) the required information is often scattered in several subtrees. These issues *do not* affect authentication – they

affect the additional information that we need to pull in order to support the teaching activities.

One option was to redesign Moodle's LDAP component, which pulls information every time a user logs into the system. However, this idea was discarded and we opted for a different solution. Since the information that we are pulling changes rarely, we decided to create two scripts:

1. The first script pulls information from the directory server and stores it in a XML file, according to a schema we have defined for this purpose.
2. Another script reads the XML file and imports the information into the Moodle database.

This approach has two benefits:

1. We isolate moodle.openfmi.net from the structure of the directory server as much as possible. Even if new changes are introduced, we will only have to rewrite the script, which pulls information. Since we expect to run it rarely (several times a year), even if a change is made to the directory server, we will have enough time to develop the required functionality (as compared to a case, where the change affects a core module, which is used for authentication of users and attempts to retrieve information on every login).
2. It will not make upgrading or migrating to a new version harder. This is not true in the case where a core module gets modified.

Currently we use other scripts to pull information from the directory server. We hope to begin the development of the new ones in the following months.

2.3 Mobile Devices Support

An increasing number of our users (both teachers and students) would like to access moodle.openfmi.net from a mobile device (both smartphones and tablets). This is an area, which we have not covered and we currently do not provide support for mobile devices. The first step in this direction will be to upgrade moodle.openfmi.net to the latest stable version of Moodle (at the time of the writing of this article it is 2.2.2), which provides improved mobile device support. This however is just the first step, as we will have to explore how mobile access to the system will change the e-learning environment and how we can adapt and bring the leading good-practices in this area to SU-FMI.

3 Conclusion

Our Moodle system is a valuable asset to SU-FMI. It provides great new opportunities for the teaching staff, the students and the administration. Even though the official Moodle installation provides useful functionality out of the box, the larger an educational organization is, the more likely it is to require additional customizations. We hope that these and any future changes that we

introduce to our system will help to create a better learning environment for everyone at SU-FMI.

4 Acknowledgment

This research is supported by contract 127/2012 of Sofia University Research Fund – 2012.

5 References

1. Moodle Course Management System. http://moodle.org/
2. http://moodle.openfmi.net/
3. Semerdzhiev, A., Trifonov, T. Nisheva, M.: Automated Tools for Assisting the Educational Process in Informatics. In: Proceedings of the International Conference on Application of Information and Communication Technology in Economy and Education (ICAICTEE) 2011, Sofia, Bulgaria, pp. 429–434. ISBN 978-954-92247-3-3 (in Bulgarian)

Choosing Approach for Data Integration

Hristo Hristov

christo_christov@yahoo.com

Abstract. The data integration is broad topic with ubiquitous presence in scientific area and enterprises. An overview of data integration field including definitions and reasons for disintegration shows that there are several approaches for data integration. For solving specific data integration task should be chosen one approach to be followed. For the purpose several criteria for evaluation of the approaches are proposed and comparison of the main integration approaches is presented. Then two methods for choosing the best approach are proposed - one with formula for calculation of sum of distances to the desired solution and another more business oriented based on radar charts.

Keywords. Data integration, approaches, definition, comparison, choosing

1. Introduction

In an organization (business, government, scientific) could coexist various information technology systems and tools with different characteristics and purposes. Each system could hold information not presented anywhere else. In the same time there could appear questions which need data from several data sources so to be answered. In many cases different data sources are not designed to work together which leads to data disintegration. Reasons that leads to disintegration could be lack of coordination between different units of the organization, merges and acquisitions, geographical separation or different level of adoption of new technology [1][27]. Also different tools for data management could coexist in the organization which are not designed to work together and in same manner. Task became even harder when a need external for the organization data sources to be used appear, which again address lack of coordination and even lack of complete information for availability and semantic of the data.

There are added benefits if the data otherwise separated and isolated could be combined and used as a whole - to be integrated. For example a company could have several databases in each company store and one more for the warehouse.

Only after using the data from all databases the company to be able to answer a question which stock is less than usual and need to be purchased or produced more items from it and exactly – how many. In order to see the scope of the data integration task in this article first we aim to try to find definition (or definitions) for information integration along with the reasons for the need of data integration. Different approaches for integration of data could be used to solve the data integration task and we are presenting here the most popular of them. But the different approaches vary very much in their characteristics and it is not easy to replace one approach with another. The question which appear naturally is which approach should be used for the specific task. Choosing the best one for the requirements of the specific case should be based on clear criteria. We propose here a set of criteria to be used for argument-based choice. We will give our proposal for evaluation grades for each approach to be used as parameters for the evaluation. At the end we will make proposal how the evaluations to be made so to be found the best choice of data integration using two methods - formula based and chart based.

The rest of the paper is organized as follows. In section 2 we will make an overview of data integration definitions, will take an outlook on the problem of heterogeneity of data sources and the essence of the theoretical base of data integration. In section 3 starting with more broad view of integration task we will list the main data integration approaches. In section 4 we will propose set of criteria for evaluation of the approaches for data integration. In section 5 we will propose evaluation of each approach on the proposed criteria. In section 6 we will summarize the evaluation of approaches and will propose grades for each approach on each criterion. Then two methods for choosing the best approach will be presented. At the end we will make some conclusions in section 7.

2. Definition of data integration

The definitions of information integration and particularly data integration are many. This is ubiquitous problem with various of forms and appliances. Data integration could be considered the situation when several data sources, each with an associated local schema, are integrated to form a single eventually virtual database with an associated global schema, transforming source data model to the global schema data model [4]. Similar definition is given in [13], where the data integration is the problem of combining data from different sources to provide the user with a unified view over their data. The end user should be provided with uniform interface to data sources but these data sources should remain independent [12]. According [16] it is meaningful information exchange among systems not originally designed to work together. Other important

characteristics of data integration is to be noninvasive as advocated in [17] where the data integration is presented as not changing the sources in a difference with schema integration where the goal is to re-engineer and re-implement the sources in a single information system. Information integration refers to the category of middleware which lets applications access data as though there were in a single database [9]. Also it could be considered as family of applications aiming in access data from two or more databases and build from it a single database with data from all sources but presented in a single unit [16].

From the business point of view the problem of the integration is commonly referred as enterprise integration and particularly for information integration there is met the term enterprise information integration [14]. Enterprise integration covers Enterprise Information Integration and Enterprise Application Integration [14].

In many definitions could be met the term heterogeneous data sources. Accordingly [10] heterogeneous sources are such sources, that have different way of representation of data and knowledge about the world. There could be different kinds of heterogeneity. [11] proposes the heterogeneity to be divided on structural and semantic one. The structural heterogeneity means that different information systems store their data in different structures and semantic heterogeneity means the meaning of the data is different in different systems. Heterogeneity could be seen in hardware and operating systems, data management software, data models or schemas and data semantics, middleware, user interfaces or in business rules and integrity constraints [14].

A data integration system appears to its users as schema presented as mediated of global schema. This schema should have defined relation with the information sources which relation could be thinked as mappings that describe the semantic relationships between the mediated schema and the schemas of the sources [3]. Two main approaches for mapping representation are known as Local-as-View (LAV) and Global-as-View (GAV). We have case of GAV mapping when the content of each element of the global schema is characterized in the terms of a view over the sources [13]. This approach is intuitive and straightforward in realization for example with predefined queries for accessing each source and metadata how to merge data from different sources. We have case of LAV mapping when the content of each source is characterized in the terms of a view over the global schema [13]. This approach is more obscured and hard to be realized, but it has some advantages as independence of the sources and easily adding new data sources to the system without impact on the definition of the existing sources. Global-Local-as-View (GLAV) is an approach which combines the expressiveness of both LAV and GAV [2].

3. Approaches for data integration

If we look more generally, the integration can have four forms [8]:

- Portals integration – bringing disparate applications together in a single point of entry (typically via Web interface)
- Business process integration – meaning orchestration of different processes
- Application integration
- Information integration –integration of different kinds of knowledge (e.g. aligning and merging domain ontologies [25, 26]) and most often data integration what the rest of this paper is about

The portals, application integration and information integration are the areas covering data integration. Data integration approaches could be divided on two main architectures - virtualization approach and materialization approach [9]. In virtualization approach data resides in its original place and the integration system access the data sources each time when it is asked to provide data as it does not have the data by itself. In the opposite - in the materialization approach the data in the source systems is accessed periodically and copied in the integration system so the queries on data are served from the copy and the source systems will not be contacted till the next iteration of data synchronization. The main examples of data integration could be distinguished are Manual (Personal information management), Portals. Integration trough applications, Mediated query systems, Data Federation, Data Warehouses, Operational data stores and Peer-to-peer. In chapter 5 we will make more detail overview of each of these approaches.

4. Criteria for comparison of the approaches

In case we are faced in front of data integration task there are several steps to be followed in order to fulfill it. For example in case of business related task such steps are data understanding, standardization, specification and execution [7]. For the time of execution there have to be taken decision which approach should be used. Once an integration approach is chosen, it is hard to switch [7] so it is preferable to base our decision on some criteria. Such criteria could be derived from the challenges which stand in front of data integration systems. Some of the challenges accordingly [10] are possibility for scalable sources integration together with automated discovery of new sources and possibility to deal with unreliable sources, how easy is configuration, management and support of the system, data protection and secure data access. Other criteria come from the need to be chosen

between LAV and GAV approaches for source schemes mapping. GAV approach is more tolerant with changes in the integrated (mediated) schema, as LAV is more suitable for rapid changes in the source schemes. Also in [7] are proposed several types of requirements to data integration task which could inspire us for our choice of criteria – qualities of service, qualities of data, physical constraints and policies. To enrich our set of criteria for evaluation, we will turn to a different area namely - the project management, as building of data integration solution usually is made under the shape of a project in some organization. As could be seen in [18] some of the most important criteria for success of a project are cost, time and quality. These criteria are named the Iron Triangle success criteria.
We are proposing the following criteria aimed to evaluate scalability of sources, need for automated discovery of new sources differentiate the need for LAV or GAV mapping:

- Sources stability - meaning how rare are the changes in the schema of the source and in that way how big is the need for automated sources discovery and how appropriate is the approach for GAV
- Integrated schema stability - meaning how rare are changes in the integration schema - and in that way will the specific approach provide convenient mechanisms for facing ever changing environment and also how appropriate is the approach for LAV

The next group of criteria are consequence from the challenges list indicated in [10]:

- Data volumes - the level of scalability and the relative size of the sources used accordingly the challenges list presented above.
- Security - what level of security could be guaranteed
- Complexity - evaluation of easiness of configuration, management and support of the system

Follows a group of criteria based on Iron Triangle criteria for success triplet:

- Budget - relative cost as money to acquire the integration system - the cost criterion from Iron Triangle triplet.
- Time for execution - how much time will be needed for development of the system

The third criteria from the Iron Triangle is quality. Different organizations can have different perception about what is the meaning of quality for data integration solution. For example except other criteria already listed, the organization can have requirement the source data to be available for queries shortly after they appear. Or the system to be easy for the users which could mean what training the users should pass so to be able to work with the system. Here could be added also the need of system administrators or in general - IT staff to support the solution. Despite these criteria are not all that can be chosen, they could be taken as criteria

giving evaluation of the quality of data integration approach. That's why we will add the following criteria to our set:

- Actuality - how short is the period between appearing of data in sources till their availability in the integration schema
- User Experience - how much trained and experienced the user should be to use the integrated system
- IT staff - what is the need of IT specialist to support the system in background

5. Evaluation of the importance levels for each criteria

The goal is to be proposed evaluation on each of the approaches accordingly these criteria. In every case when a data integration approach need to be chosen, we should give evaluation of the desired solution on the same set of criteria and to be provided an evaluation methodology so the best approach to be chosen. Below we will pass trough the approaches presented in section 3 and will propose levels for each of the criteria.

Manual

In this form of integration the main role in data integration plays the user itself. He or she is using various tools and different query languages to access, extract, manage and store the data. The user should be familiar with data as destination and semantics and to be granted with respective security rights [14]. The area of Personal information management (PIM) covers the manual approach. PIM refers to both the practice and the study of the activities a person performs in order to acquire or create, store, organize, maintain, retrieve, use and distribute the information needed to meet life's many goals [6].
Manual approach could serve well if the task for integration is small, the user is experienced (or quickly can become) and has access to all data needed. But this approach is very basic and inefficient if we are talking for large amounts of data. Together with entirely manual data integration, some works proposed solutions for computer added integration without reaching the magnitude of mediators or data warehouses. For example in [5] is presented information integration architecture that stands between search engines and traditional information-integration architecture. Manual data integration is flexible approach and the execution of the integration steps could be adapted relatively easier if changes in the sources appear but at the other side it could take time and effort so the importance of source stability is considered to have medium importance. The integration schema often is in the user mind only so we can consider that its importance is low. Collection

of data from sources can take a lot of time during which the data to become obsolete so data actuality is low. Data volumes cannot be very big having in mind it is manual integration from one or several persons and the volumes are limited from what they can collect and manage to manipulate. User experience should be high in order to understand data and to be able to use the tools for integration. There is low need from IT staff, usually the security is not guaranteed because the manual approach and techniques for data collecting. Budget needed is low. Time for setup the tools for integration usually is relatively short. Complexity of the approach is high because usually the user does not bother to document the integration approach so take over on his or her work is difficult task.

Portals

Common user interfaces and Portals as presented in [14] looks to be step to reduce the disadvantage of arbitrary choice of different query tools and languages to be used and hiding the details for data extraction, storage and security. Portals provide a means of collecting information—perhaps from quite different data sources —and putting the results together for the user to see in conjunction. Portals could add some extra functionality like tailoring to the users needs accordingly to the activity history. Even Portals could provide some tools for creation of own applications [1]. But this approaches in general limit the possibility for data manipulation for the user and if the data on the interface is not in the needed format, not aggregated, not filtered good enough, actually the task became equal to manual integration approach.
Adding new sources in portals and providing the ability the user to see the data on the screen rarely are very dynamic but some possibilities exist in that direction so the importance of sources and integration schema stability could be evaluated as moderate. Portals usually give the actual data found in the sources so the actuality level is high. Data volumes are bigger than those for the manual integration and are equivalent to the volumes possible for management trough application - so the level for data volumes is evaluated to medium. Usually the portal interfaces are very user friendly so there is no need of special user experience. The IT staff should be trained to support the portal itself and the applications for accessing it (browsers in most cases) and the importance of their robustness is medium. High security level could be achieved. Budget, time for execution and complexity of the solution does not pretend for highest levels so they are considered to have medium importance.

Application integration

The application integration includes communication of applications, doing similar or complementary operations and integration of services so to be provided single service access point [14]. These kind of approaches are based on special-purpose applications or restrictions in the environment for information management. The applications access sources of interest directly and combine the data retrieved from those sources with the application itself. This approach always works, but it is expensive (in terms of both time and skills), fragile (changes to the underlying sources may all too easily break the application), and hard to extend (a new data source requires new code to be written) [1].
Applications require the sources to be stable and the intended integration schema not to change. There is possibility for online data collection so results actuality is high. But usually applications cannot handle very large volumes of data especially if we are talking for non-distributed computing. The user should be experienced with the application but the required level could be reduced with proper GUI creation. IT staff for maintenance should be moderately experienced if the application is error free and convenient for administration. High security levels can be achieved. Having in mind that this is specific application developed especially for given set of sources and integration needs, the budget, time for execution and complexity of the system usually are high.

Mediated query systems

Mediator is a software component that supports a virtual database. The system stores no data of its own, but translates the user's query in queries to the sources. After the answer from the sources is received, the mediator integrates the results and returns to the user an answer [15].
Mediators can work in situation of high sources and integration schema instability so importance of these criteria is evaluated to low. Mediators access fresh data so the actuality is high. Data volumes that could be accessed could be with high degree. The user experience should be same as for the application approach - the user should be able to work with the integration software. The IT staff should be highly experienced so to be aware on the data sources in one side and mediator's software in another so to maintain correct mappings and workability of the system. High security level could be achieved. The cost of the solution could be accepted as moderate compared for example with data warehouse where at least hardware requirements could be substantial. Time for execution also is lower than needed for data warehouse so the level is medium. But complexity of the system exceeds that of the Data Warehouse and is high.

Data Federation

In this approach sources are independent but one source can call on others to supply information [15]. The system transforms user's queries into specific requests for data to the underlying data sources in a way transparent to the user [1]. Some specific for the data source functions could be used in that way. Each connection between two data sources is specific for the calling and called databases and if the databases are many the connections that need to be established will be many as well.

Adding new sources to the federated database could be challenging task especially the new sources are not compliant with the other sources and the fact that all existing database nodes should be managed to connect to the new database. But the work could be limited in the interfaces between databases only or connection to the newly added source could not be needed for all existing databases and the main functionality could need little changes. That's why the impact of the sources and integrated schema stability could be accepted as moderate comparing with application approach for example. Online access to data is used to construct the answer to the query and the actuality level here is high. Data volumes could be significant and this criteria is with high degree. The system hides the process of accessing the sources in background and choosing the exact execution plan so the user has to know only to work with the interface which could be said to be medium level importance. IT staff supporting the data warehouse should be highly experienced. High level of security could be achieved. Having in mind distributed characteristics of the databases there could be achieved some budget savings so the importance of this criteria is medium. Time for execution could be less than those for data warehouse approach if for example the databases already exist. But complexity of defining algorithms for query execution plan choosing and query execution leads the complexity criteria to be with high degree.

Data Warehouses

It is common data storage keeping copies of the data from the source systems. Data is extracted, transformed in appropriate way and loaded in the data warehouse system which is known as Extract-Transform-Load mechanism. During the transformation phase the data is cleansed, standardized, de-duplicated if needed and filtered. Many issues arise at this stage, concerning the data integrity [21] and different methods can be used to achieve better results [24]. Some processing may be applied further as aggregations or more complex logic algorithms applying. Often the resulting schema has special design helping queries for analysis and decision taking. Data Warehouse is intended to serve queries which use large

volumes, many rows at a time and seldom is interesting from a single event despite it is possible. Often the data in the Data Warehouse is updated periodically (for example each night) and the system is not intended to present the most accurate state, but some snapshot taken some time ago. Changes in the underlying sources may cause changes to the load process, but the part of the application that deals with data analysis is protected. New data sources may introduce changes to the schema, requiring that a new load process for the new data be defined. SQL views can further protect the application from such evolutions [1].
Data Warehouse relies on ETL and staging area to take data from sources and to store it in own schema. That's why there is some layer which reduces the impact from changes of the sources and medium importance of sources stability could be accepted. Integrated schema stability could be considered as high level of importance and some techniques could be applied for performance optimization[28]. The data in Data Warehouse is usually not actual but present the state from some passed point in time. But the refresh of data is regular and in some cases is near to the online access so we can accept middle level of importance for the data actuality. Data volumes could be considered with high level - usually the volumes of data stored is huge. User experience should be higher compared with that of portals for example but lower than manual approach so medium level is proposed here. The IT staff supporting the data warehouse should be highly experienced. High level of security could be achieved. Although DW2.0 paradigm is established now [23] and different optimizations are proposed [22] this approach is expensive and usually the budget should be significant. Time for development also is high. But the complexity of the solution is less than that for the mediators and could be accepted as medium.

Operational data stores

It is again common data storage like data warehouse. But the difference is that here the updates in the source systems are propagated immediately in the data store. But the data is not cleansed, aggregated and no history is supported [14].
Because it is too close to data warehouse approach, the criteria importance for data stores will be evaluated with comparing with data warehouses. Sources stability, integrated schema stability, actuality and data volumes are considered same - respectively medium and high. User should be more experienced, because he or she needs to do some data cleansing, to know the semantics of the data coming directly from sources. But for the same reason the need of experienced IT staff is reduced and medium levels for this criteria is accepted. Security level also is evaluated as medium as no special preliminary data processing is made. The budget and time for execution are reduced to medium. Complexity is also less than those of data warehouse and is evaluated to low.

Peer-to-Peer

Peer-to-Peer is a system architecture constituted by a number of peers. Each peer holds local data and is connected to other peers. Every peer acts as both client and server and provides part of the overall information available from a distributed environment. [19]. Especially for the purpose of data management and data integration is used Peer Data Management System (PDMS) [20]. Queries are posed to one peer and data is looked both internally in this peer and externally in the other connected peers or in peers connected trough these peers and so on.
For Peer-to-Peer data integration the sources are ever changing peers so the sources stability has low importance. The integration schema also can be changed often mainly with the newly added or disappeared sources. Actuality of the data is high because each time a query is posed we receive the most current state available. The question for data volumes has two sides. There is access to potentially very large volume of data especially if the P2P uses actually Internet as connecting environment between the peers. But in the same time this same connection environment does not allow big data transfer between peers which to reach the peer which is asked for answer of the query. User experience is low to moderate need - in most cases the software managing peers will hide complicated details from the end user. The need of IT staff is low - actually low experienced users can be able to install and deal with the software. Security is one of the open questions for researchers in this area so we consider the value to be low. Development of the management software could be considered relatively low compared with the volume and count of the peers which could be included in the system. In that direction also the budget and time for execution could be considered to have low importance. Complexity also could be considered low as each peer will need to know its own functionality and just to be aware for the schema of the nearby peers.

6. Comparison of the approaches

For each criterion we set a number between 1 and 3, where 1 means low grade, 2 - medium grade and 3 - high grade. Accordingly the evaluation from the previous point we receive the following table of levels:

Table 1. Grades of the approaches for each criterion

	Manual	App	Portals	Mediators	Federation	DWH	Data Store	P2P
Sources stability	2	3	2	1	2	2	2	1
Int. schema stability	1	3	2	1	2	3	3	1
Actuality	1	3	3	3	3	2	2	3
Volumes	1	2	2	3	3	3	3	2
User Exp.	3	2	1	2	2	2	3	2
IT staff	1	2	2	3	3	3	2	1
Security	1	3	3	3	3	3	2	1
Budget	1	3	2	2	2	3	2	1
Exec. Time	1	3	2	3	2	3	2	1
Com-plexity	3	3	2	3	2	3	1	1

We will propose two methods for choosing the approach which best fit our needs. In both cases we have to have in advance our decision what grades should have our solution so to try to find the approach which is closest to our desires.
The first method is based on the sum of differences between criteria grades for each approach compared with the desired solution. Let A is the set of approaches and A= {Manual, App, Portals. Mediators. Federation, DWH, Data Store, P2P}. Let C is the set of criteria and C = {Sources stability, Int schema stability, Actuality, Data volumes, User Experience, IT staff, Security, Budget, Time for execution, Complexity}. Let G_{ac} is the grade of approach a for criteria c. Let S_c is the desired level for the solution we need for the criteria c. We use the following formula:

$$D_a = \sum (S_c - G_{ac}) \text{ for each } a \in A \qquad \textbf{(1)}$$

The best choice is the approach a for which D_a is minimal.

The Second proposed method is to be used radar charts (or somewhere named net charts). The following diagrams presents one radar chart for each of the approaches from Table 1:

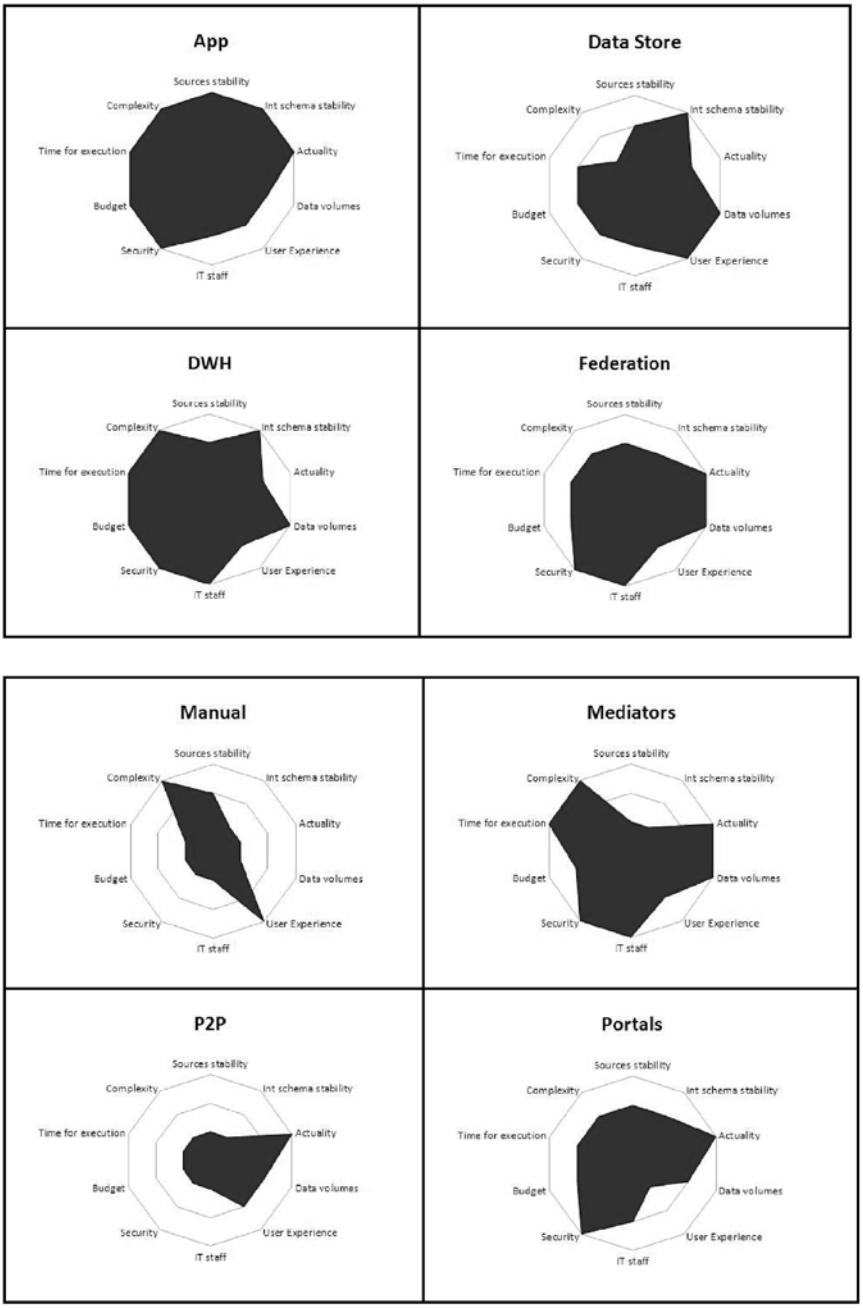


Fig. 1. Radar Charts for each of the approaches

To choice of the most desired solution is the one which chart mostly fits our expectations how the solution radar chart should look like.

7. Conclusion

In this work were presented definitions and the main approaches for data integration. Set of criteria for evaluation of the different approaches were presented and for each approach was given evaluation of the degree each criteria applies to the approach. As result we received a table with grades values which can be used for comparison of different data integration approaches. Two methods to choose the best approach for the case are proposed. First is based on sum of the differences between the grade of each approach accordingly each criteria with the desired grade for that approach and then choosing this approach, which have minimal difference. The second is more visual. We use radar chart to present the charts for each approach. The choice of the best approach should be made accordingly the outlook which best fit the needs of the organization.

The proposed methods could be adapted to the needs of the organization. The evaluation of the approaches could be changed so to best fit the knowledge and experience of the evaluating team. Also more data integration approaches could be added - variants of the existing or pure new approaches with different characteristics.

Acknowledgment. This paper is supported by Sofia University "St. Kliment Ohridski" SRF under Contract 134/2012.

References

1. Haas, L.M., E. Tien Lin, M. Tork Roth: Data integration through database federation. In: IBM Systems Journal, 41:4, pp. 578-596 (2002)
2. Friedman, M., Levy, A., Millstein, T.: Navigational plans for data integration. In: Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence (AAAI '99/IAAI '99). American Association for Artificial Intelligence, Menlo Park, CA, USA, 67-73 (1999)
3. Halevy, A., Rajaraman, A., Ordille, J.: Data integration: the teenage years. In: VLDB'2006: Proceedings of the 32nd international conference on Very large data bases.VLDB Endowment, pp. 9-16 (2006)

4. Williams, D., Poulovassilis, A.: Combining Data Integration with Natural Language Technology for the Semantic Web. In: Proceedings of Proc. Workshop on Human Language Technology for the Semantic Web and Web Services, at ISWC'03 (2003)
5. Salles, M.: Pay-as-you-go information integration in personal and social dataspaces. Diss. ETH. NO 18079, ETH Zurich (2008)
6. Jones, W.: Keeping Found Things Found: the Study and Practice of Personal Information Management. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2007)
7. Haas, L. M.: Beauty and the beast: The theory and practice of information integration. In: ICDT, pages 28-43 (2007)
8. Jhingran, A. D., Mattos, N., Pirahesh, H.: Information integration: A research agenda. IBM Syst. J. 41, 4, 555-562 (2002)
9. Mohania, M., Bhide, M.: New trends in information integration. In: Proceedings of the 2nd international conference on Ubiquitous information management and communication (ICUIMC '08). ACM, New York, NY, USA, 74-81(2008)
10. Halevy, A., Li, Ch.: Information Integration Research: Summary of NSF IDM Workshop Breakout Session (2003)
11. Wache, H., Vцgele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hьbner, S.: Ontology-based Integration of Information - A Survey of Existing Approaches. In: Proceedings of IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA,, Vol. pp. 108-117 (2001)
12. Bennett, T., Bayrak. C.: Bridging the data integration gap: from theory to implementation. SIGSOFT Softw. Eng. Notes 36, 4 (2011)
13. Lenzerini, M.: Data integration: A theoretical perspective. Tutorial at PODS 2002 Madison, Wisconsin, USA (2002)
14. Ziegler, P., Dittrich, K.: Three Decades of Data Integration - All Problems Solved?. In: 18th IFIP World Computer Congress (WCC), volume 12, Building the Information Society, pages 3-12, Toulouse, France (2004)
15. Ullman, J. D., Garcia-Molina, H., Widom, J.: Database Systems: The Complete Book (1st ed.). Prentice Hall PTR, Upper Saddle River, NJ, USA (2001)
16. Seligman, L., Rosenthal, A., Lehner, P., Smith, A: Data integration: Where does the time go? IEEE Data Engineering Bulletin (2002)
17. Koch, C.: Data Integration against Multiple Evolving Autonomous Schemata. PhD thesis, Technische Universitat Wien, Austria (2001)
18. Atkinson, R.: Project management: cost, time and quality, two best guesses and a phenomenon, its time to accept other success criteria. International Journal of Project Management Vol. 17, No. 6, pp. 337-342 (1999)
19. Lenzerini, M.: Principles of peer to peer data integration. DIWeb2004. Available from http://www.doc.ic.ac.uk/~pjm/diweb2004/DIWeb2004_Part2.pdf (visited May 12, 2012) (2004)
20. Halevy, Y., Ives, G., Suciu, D., Tatarinov, I.: Schema mediation for large-scale semantic data sharing. The VLDB Journal 14, 1 (March 2005), 68-83 (2005)
21. Kaloyanova K., Improving Data Integration for Data Warehouse: A Data Mining Approach, Proceedings of the International Workshop "COMPUTER SCIENCE AND EDICATION", Borovetz-Sofia, Bulgaria, June, 2005, ISBN 954-535-401-1, pp 39-44
22. Ina Naydenova, K. Kaloyanova, "Some Extensions to the Multidimensional Data Model, ISGT'06: IEEE 2006 John Vincent Atanasoff International Symposium on Modern Computing, October 2006, Sofia, Bulgaria
23. Hristov N., K. Kaloyanova, Applying a Seven Stream Approach for Building a DW 2.0, Proceedings of the Fourth International Conference on Information Systems & Grid Technologies, 28 - 29 May 2010, Sofia, Bulgaria, 92-101

24. Kovacheva Zl., Application of Neural Networks to Data Mining, Sultan Qaboos University Journal for Science, Muscat, Oman, Vol.12, Part 2, December 2007, pp. 121 – 141.
25. Zlatareva, N., Nisheva, M.: Alignment of Heterogeneous Ontologies: A Practical Approach to Testing for Similarities and Discrepancies. In: D. Wilson, H. Chad Lane (Eds.), Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference (Coconut Grove, Florida. May 15–17, 2008). ISBN 978-1-57735-365-2, AAAI Press, Menlo Park, California, 2008, pp. 365–370
26. Nisheva-Pavlova, M.: Mapping and Merging Domain Ontologies in Digital Library Systems. Proceedings of the Fifth International Conference on Information Systems and Grid Technologies (Sofia, May 27-28, 2011), ISSN 1314-4855, Sofia, St. Kliment Ohridski University Press, 2011, pp. 107-113
27. Savoska S, B. Dimeski, "The Future Challenge: Data Warehouse As a Method of Data Integration from Public Administration Legacy Systems for Decision-making in Macedonia", Fifth INTERNATIONAL CONFERENCE, INFORMATION SYSTEMS & GRID TECHNOLOGIES, May 27 - 28 2011, Sofia, Bulgaria
28 Dimovski A., G. Velinov, D. Sahpaski, Horizontal Partitioning by Predicate Abstraction and Its Application to Data Warehouse Design, Advances in Databases and Information Systems - 14th East European Conference, ADBIS 2010, Novi Sad, Serbia, September 20-24, 2010, Proceedings. Volume 6295 of Lecture Notes in Computer Science, pages 164-175, Springer, 2010

Mining Bibliographic Data

Tsvetanka Georgieva-Trifonova
Department of Mathematics and Informatics
"St. Cyril and St. Methodius" University of Veliko Tarnovo, Veliko Tarnovo, Bulgaria
cv.georgieva@uni-vt.bg

Abstract. In this paper, the system *bgMath/DM* for mining bibliographic data is proposed. The implemented system can be useful for the users maintaining their electronic libraries with publications in order to monitoring, evaluating and comparing the scientific development of particular researchers, entire research groups, certain scientific fields and problems.

Keywords: data mining; bibliographic information system; electronic library

1 Introduction

In the last years, the digitalized form is important part of the creation, the distribution and the usage of the scientific literature. This fact concerns the periodical issues, as well as the conference proceedings and even the monographs and the reference books. The distribution and publishing the materials in Internet expedites the development of a large number of bibliographic systems integrated with search engines.

In [5], the system *Personal eLibrary bgMath* is represented, whose aim is different: easy manipulation with these materials (papers, dissertations, reports, etc.) and data about them, that are used repeatedly and are necessary at every turn in the scientific activity – scientific researching, writhing papers and dissertations, preparing reports and Web pages or application documentation. The bibliographic system for scientific literature *bgMath*, utilized from one or more users working at one or more scientific sections, allows accumulating data which are of interest for analyzing. This is the basic motivation for applying the algorithms for data mining on the bibliographic data obtained from it.

The main purpose of the system *bgMath/DM* is to provide a possibility for monitoring, evaluating and comparing the scientific development of particular researchers, entire research groups, separate scientific areas and problems.

More concretely, the implemented system can be utilized for the following:

- analyzing the bibliographic data collected from the usage of *bgMath* from one or group of researchers by applying association analysis and clustering;
- establishing useful relationships between the attributes characterizing the

publications – types of the publications; journals in which the papers are published; conferences where the papers are represented; keywords; year of publishing; number of the publication's authors; number of the citations;
- grouping the records for the publications according to different their characteristics.

The basic features of the system *bgMath/DM* are divided by four groups:
- loading the data in the data warehouse periodically by a given schedule;
- performing the algorithms for data mining on the data from the warehouse;
- browsing the results from applying the algorithms for data mining through Microsoft Excel application;
- exporting the summarized data in PDF, HTML, XML, others formats.

2 Data Mining Systems and Bibliographic Databases

OLAP (*online analytical processing*) technology and the algorithms for data mining could be applied to support solving important problems regarding the bibliographic databases in the libraries.

OLAP systems can be used for periodic reporting and data integrity checking. Analysts can interactively browse hierarchical and summarized data in order to extract new knowledge from the database. The traditional relational systems for database management that does not support OLAP, are appropriate for storing the data needed for daily activities and transactions processing. They are not suitable for performing complex queries that access large datasets and make multiple scans, joins and summaries, because they require more time for answer [3, 12]. Minimizing the response time of these queries proves crucial influence at designing OLAP applications.

Recently, OLAP systems on bibliographic databases are implemented. In [10], OLAP system for analyzing data in the Slovenian national bibliographic database *Biomedicina Slovenica* is proposed. DBPubs [2] is a system for analyzing and exploring the content of database publications by combining keyword search with OLAP operations. The purpose of the system *bgMath/OLAP* [9] is warehousing and online analytical processing bibliographic data.

Data mining systems are implemented in which the algorithms are applied on bibliographic databases in order to searching and browsing bibliographic data, extracting useful information from them, discovering communities of researchers. [13] considers the implementation and the performance of the developed bibliographic system with applying the algorithms for text mining. In [14], a user interface for searching, browsing and mining bibliographic data is described.

In [11], a community mining system using bibliography data is proposed, in order to find communities of researchers. DBconnect [23] is a prototype that

exploits the social network coded within the DBLP database by drawing on a new random walk approach to reveal interesting knowledge about the research community and even recommend collaborations. In [19] the authors propose a formal definition to consider the similarity and dissimilarity between individuals of a social network and how a graph-based clustering method can extract research communities from the DBLP database.

In this paper, the system *bgMath/DM* is represented, whose purpose is applying data mining algorithms to exploring the data obtained from the bibliographic system for scientific literature *bgMath*. The developed system allows analyzing the number of the publications and the citations by years, by keywords, by authors, by scientific sections, etc.

3 Designing and Implementing the System *bgMath/DM*

The development of the system *bgMath/DM* includes designing and implementing a data warehouse; a package for loading the data in the warehouse; applying algorithms for data mining; a client application for visualizing the results.

3.1 Architecture of the System *bgMath/DM*

The architecture of the system *bgMath/DM* is represented in figure 1.

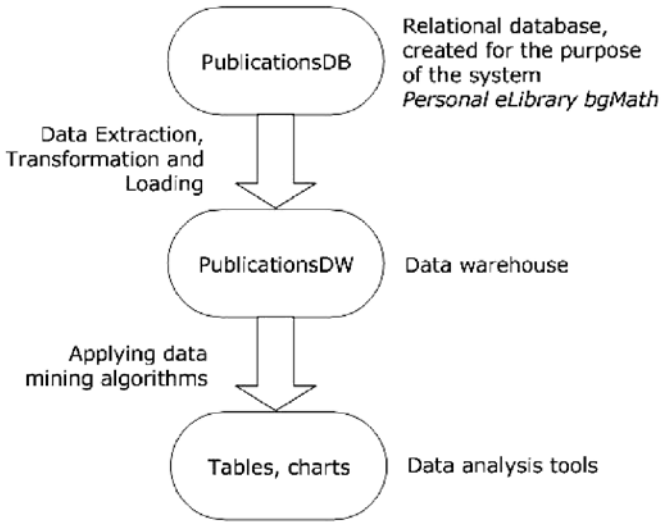


Fig. 1 Architecture of the system *bgMath/DM*

The relational database `PublicationsDB` is designed and created for the purposes of the system *Personal eLibrary bgMath*. The structure of this database is described in detail in [5].

Before applying algorithms for data mining it is necessary to perform data cleaning, integrating, extracting and transforming. With the existing relational database system and data warehouses this previous processing could be accomplished by building a data warehouse and executing some OLAP operations on the created data warehouse. Therefore the development of the system *bgMath/DM* includes creation and maintenance of the data warehouse `PublicationsDW`. The dimension tables in the data warehouse `PublicationsDW` store the data about authors; types of the publications; journals in which papers are published; conferences where papers are represented; keywords; year of publishing; number of the publication's authors. The fact table *Papers_fact* includes attributes which refer the dimension tables and the measure attributes: *CountOfPapers* – the number of the publications, *CountOfCitationsPlus* – the number of the citations; *CountOfCitations* – the number of the citations without the self-citations. The structure of the data warehouse `PublicationsDW` and the implementation of the process of data extraction, transformation and loading (ETL) are represented in detail in [9].

Microsoft SQL Server 2008 [4, 7, 8, 15, 17, 18, 20, 21, 22] is the database management system, utilized in this research.

The data mining algorithms are applied on the data collected in the data warehouse and they are described in section 3.2.

The execution of the data mining algorithms for analyzing the bibliographic data is performed with an application that is implemented for the purpose with the means of Microsoft Excel. In section 3.3.2, exemplary results are shown; they are visualized in view which is convenient for their interpretation.

3.2 Algorithms for Mining Bibliographic Data

Data mining aims discovering useful summarizations of the data. In the present research, algorithms for mining association rules and clustering are applied.

3.2.1 Mining Association Rules

Association analysis explores the frequency of the items occurring together in a transaction database and establishes which itemsets are *frequent*.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of the considered *items*. In the association analysis, the itemsets $T = \{i_j \mid i_j \in I\}$, that are subsets of I, stored in the database and liable to analysis, are called *transactions*. Let $D = \{T_1, T_2, \ldots, T_m\}$ be the set of transactions whose information is accessible for analysis. The set of transactions

which contain the itemset F, we denote by $D_F = \{T_l \mid F \subseteq T_l;\ l = 1, \ldots, m\}$, where $D_F \subseteq D$.

The proportion of the number of the transactions which contain F to the total number of the transactions is called *support* of the itemset F and is denoted by $supp(F)$, consequently

$$supp(F) = \frac{|D_F|}{|D|}.$$

Discovering the association rules requires setting a minimal value of the support of the searched itemsets $supp_{min}$. We say that a given itemset F is *frequent itemset*, if the value of its support is greater than the chosen minimal support, i.e. $supp(F) > supp_{min}$.

An association rule is an implication of the form $X \rightarrow Y$, where $X, Y \subset I$ are sets of items with $X \cap Y = \varnothing$. The set X is called an antecedent, and Y – a consequent. There are two parameters associated with a rule – support and confidence.

The support s of the association rule $X \rightarrow Y$ is the proportion (in percentages) of the number of the transactions in D, which contain $X \cup Y$ to the total number of the transactions, i.e. $supp(X \rightarrow Y) = supp(X \cup Y) = \frac{|D_{X \cup Y}|}{|D|}$.

The support is an important measure for determining the association rules because a rule that has very low support may occur simple by chance.

The confidence $conf(X \rightarrow Y)$ of the association rule $X \rightarrow Y$ is the proportion (in percentages) of the number of the transactions in D, which contain $X \cup Y$ to the number of the transactions, which contain X, i.e. $conf(X \rightarrow Y) = \frac{|D_{X \cup Y}|}{|D_X|} = \frac{supp(X \cup Y)}{supp(X)}$.

The confidence measures the reliability of the inference made by a rule. For a given rule $X \rightarrow Y$, the higher value of the confidence means that the more likely it is for Y to be present in transactions that contain X.

The task of association rules mining is to generate all association rules which have values of the parameters support and confidence, exceeding the previously given respectively minimal support $supp_{min}$ and minimal confidence $conf_{min}$.

Discovering all frequent itemsets requires large number of computations and time. In general, a dataset that contains k items can potentially generate up to $2^k - 1$ frequent itemsets. Therefore it is necessary reducing the itemsets for which the value of the support is computed and verification is performed whether it exceeds the minimal.

The Apriori algorithm [1], proposed in 1994 year from Rakesh Agrawal and

Ramakrishnan Srikant, is based on the apriori property: each nonempty subset of a frequent itemset is also frequent itemset.

After finding all frequent itemsets, they are utilized for the generation of the searched association rules. For each frequent itemset X, the confidence of all rules of the form $X \setminus Y \rightarrow Y$, $Y \subset X$, $Y \neq \emptyset$ is verified and if a rule does not reach the minimal value of the confidence $conf_{min}$, it is removed. Consequently, if the supports of the subsets of X are known, it is possible to compute the confidence of each rule. After discovering the frequent itemsets, the supports of all subsets of X are also known, because the apriori property is satisfied.

The Apriori algorithm determines the frequent itemsets by several stages: on the first step the algorithm finds the itemset L_1 of the one-element frequent itemsets; then the itemset C_k of k-element candidate itemsets is generated for $k \geq 2$ by joining the frequent $(k-1)$-element itemsets L_{k-1} and removing these itemsets that have $(k-1)$-element subset that is not frequent; on the next step, the set of the frequent k-element itemsets L_k is generated on the basis of C_k, by verifying the support of the corresponding candidate itemset from C_k for exceeding the minimal $supp_{min}$. On each iteration k is increased with 1 ($k = k + 1$) and this process continues while $L_k = \emptyset$ for some k. The result from the execution of the algorithm is the set L, which represents a union from the sets L_k for each k.

To select interesting association rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The lift importance is one of these measures defined in [6]. It provides additional information about the found association rules. The lift of a rule is defined as:

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{supp(Y)} = \frac{supp(X \cup Y)}{supp(X)\,supp(Y)}.$$

The value of the lift importance larger than 1.0 indicates that transactions containing X tend to contain Y more often than transactions that do not contain X.

3.2.2 Clustering

Cluster analysis divides objects into groups on the basis of the information found in the data that describes the objects and their relationships. The aim is that the objects within a group (cluster) be similar to one another and different from the objects in other groups.

The K-means algorithm [16] associates a point with this cluster whose centroid is nearest. We first choose K initial centroids, where K is a user-specified parameter and determines the number of the clusters desired. Each point is then assigned to the closest centroid. The centroid of each cluster is then updated based on the points assigned to the cluster. We repeat the assignment and update steps until no

point changes clusters, i.e. until the centroids remain the same.

The formal description of the K-means algorithm consists of the following steps:

1. Select K points as initials centroids.
2. Form K clusters by assigning each point to its closest centroid.
3. Recompute the centroid of each cluster.
4. If the centroids do not change, the algorithm finishes its execution, else the algorithm returns into step 2.

To assign a point to the closest cluster, it is necessary a measure be applied for determining the closeness between the considered data. The distance between the objects supposes their representation as points in the m-dimensional space R^m. Let $X_Q \subseteq R^m$ be a set of data with elements $x_i = (x_{i1}, \ldots, x_{im}) \in X_Q$, $i = 1, \ldots, Q$.

The common measures are:

- Euclidian distance

$$d_2(x_i, x_j) = \sqrt{\sum_{t=1}^{m} (x_{it} - x_{jt})^2}$$

- Hamming distance

$$d_H(x_i, x_j) = \sum_{t=1}^{m} | x_{it} - x_{jt} |$$

- Chebyshev distance

$$d_\infty(x_i, x_j) = \max_{1 \le t \le m} | x_{it} - x_{jt} |$$

The implementation [18] of the K-means algorithm can be conform to discrete attributes by using probabilities. For the purpose the algorithm computes the probability $P(a_k, C_i)$ that the attribute A has value a_k for each cluster. Then the distance between point x_i with value a_k for the attribute A and the cluster C_i is $1 - P(a_k, C_i)$.

3.3 Client Application for Mining Bibliographic Data

For the end user, an application is implemented with the means of Microsoft Excel [15]. This application allows applying data mining algorithms and representing the results in tabular and graphical views. To implement *bgMath/DM* we utilize the algorithms Microsoft Association, Microsoft Clustering, which are based on the Apriori algorithm and the K-means algorithm described in sections 3.2.1 and 3.2.2.

The users can access multiple reports with the application and some of them

are the association and clustering analysis of:

- the type of the publications, their year of publishing and the number of the authors of the separate publications (fig. 2, 3, 4);

  Each of the reports for association analysis includes outputting the found association rules (fig. 2), the found frequent itemsets (fig. 3) and the diagram of the dependency network (fig. 4).
- the number of citations, the type of the publications and the number of the authors of the publications;
- the keywords, the number of the citations;
- the type of the publications and the number of the authors (fig. 5);
- the institutions of the authors, the type of the publications and the number of the citations.

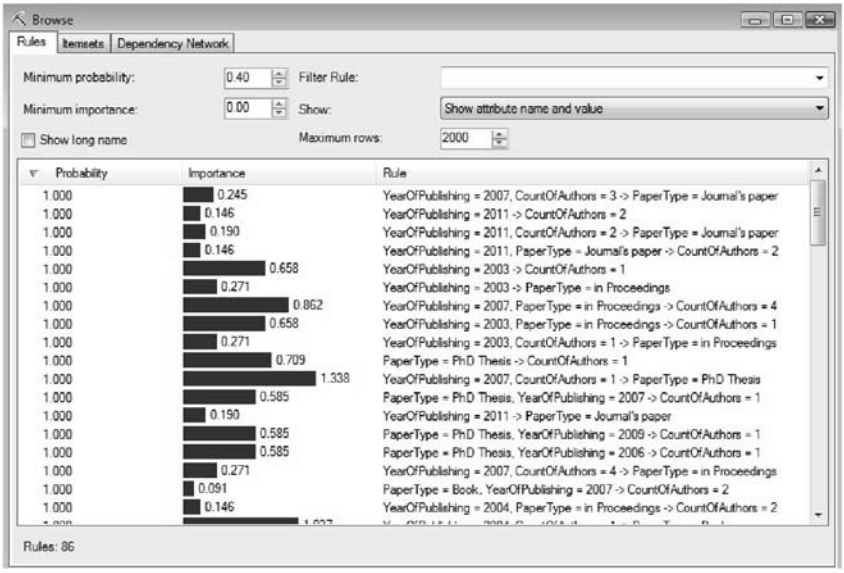


Fig. 2 Association rules

For example, the rule {*YearOfPublishing*=2007, *CountOfAuthors*=3} → {*PaperType* = Journal's paper} with values of the support 0.04 and the confidence 1.00, means that 100% of all publications published in 2007 and written by three authors, are journal's papers; as well as the journal's publications in 2007 with three authors represent 4% of all publications included in the study.

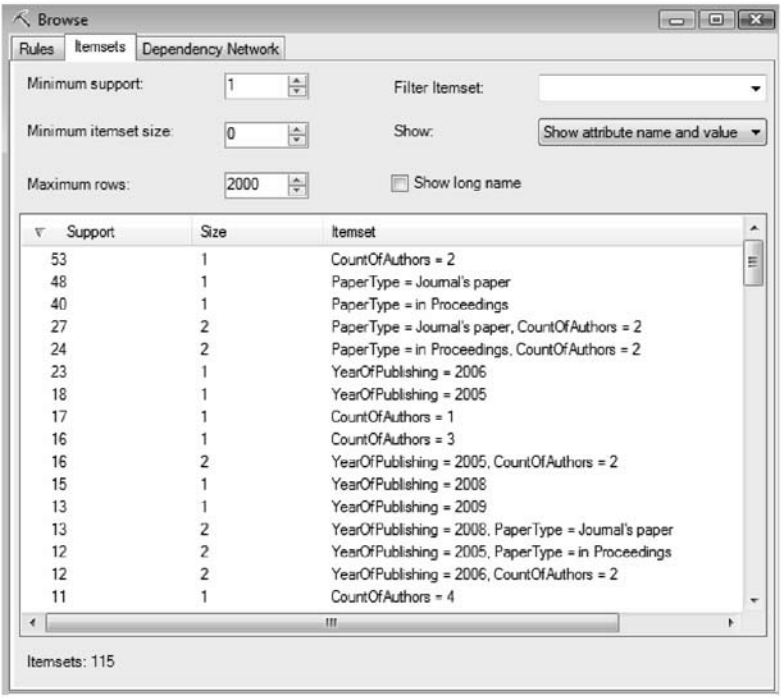


Fig. 3 Frequent itemsets

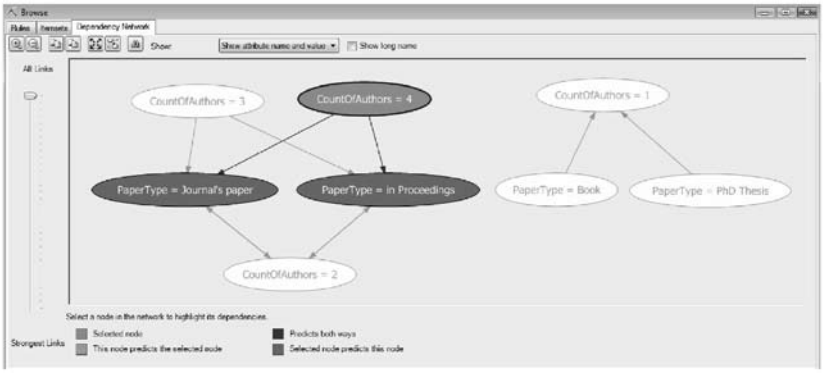


Fig. 4 Dependency network

Figure 4 shows the dependency networks between the number of the authors of the publications and their types. From the diagram we can conclude that in the most cases the publications with more than one author are journal's papers or conference's papers. Figure 5 confirms this observation.

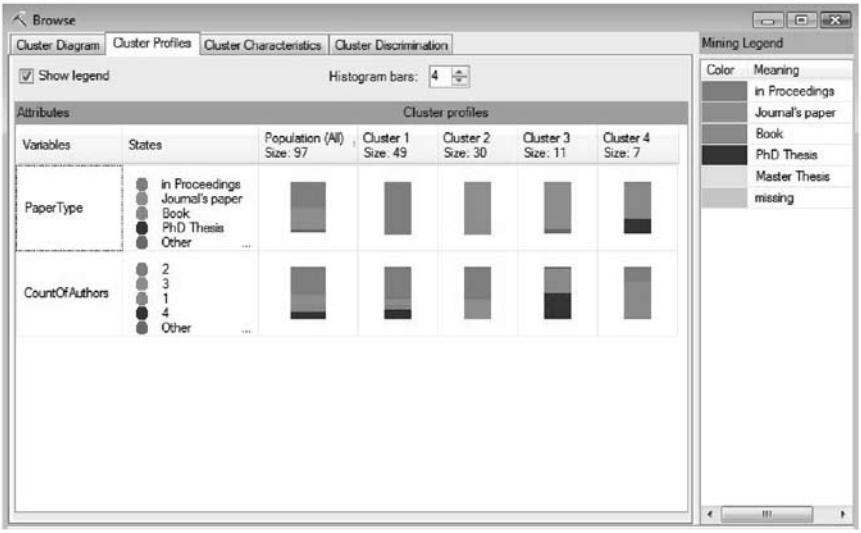


Fig. 5 Clustering analysis of the type of the publications and the number of the authors

4 Conclusion

In the present paper, an application of data mining algorithms for analyzing the bibliographic data is represented. The architecture of the implemented system for mining bibliographic data is described, as well as the utilized algorithms and the features of the application designed for the end user.

Our future work includes development of an application for data mining in the text of the publications (*text mining*), which allows performing the analysis of the different words from their contents.

Acknowledgments. This research is partially supported by the project "Modern tendencies in development of the software technologies and algorithms, the data processing and the adequate preparation of specialists in these areas", "St. Cyril and St. Methodius" University of Veliko Tarnovo, 2011.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487-499 (1994)
2. Baid, A., Balmin, A., Hwang, H., Nijkamp, E., Rao, J., Reinwald, B., Simitsis, A., Sismanis Y., Ham, F.: DBPubs: Multidimensional Exploration of Database Publications. In Proceedings of the 34th International Conference on Very Large Data Bases (VLDB '08), pp. 1456-1459 (2008)

3. Barsegyan, A. A., Kupriyanov, M. S., Stepanenko, V. V., Holod, I. I.: Technologies for data analysis: Data Mining, Visual Mining, Text Mining, OLAP. BHV-Peterburg (2008)
4. Ben-Gan, I., Kollar, L., Sarka, D., Kass, S.: Inside Microsoft® SQL Server® 2008: T-SQL Querying. Microsoft Press (2009)
5. Bouyukliev, I., Georgieva-Trifonova, T.: Development of a personal bibliographic information system. The Electronic Library. 31 (2), to appear (2013)
6. Brin, S., Motwani, R., Ullman, J. D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In Proceedings ACM SIGMOD International Conference on Management of Data, pp. 255-264 (1997)
7. Coles, M.: Pro T-SQL 2008 Programmer's Guide. Apress (2008)
8. Davidson, L., Kline, K., Klein, S., Windisch, K.: Pro SQL Server 2008 Relational Database Design and Implementation. Apress (2008)
9. Georgieva-Trifonova, T.: Warehousing and OLAP analysis of bibliographic data. Intelligent Information Management. 3(5), pp. 190-197 (2011)
10. Hudomalj, E., Vidmar, G.: OLAP and bibliographic databases, Scientometrics. 58(3), pp. 609-622 (2003)
11. Ichise, R., Takeda, H., Ueyama, K.: Community Mining Tool Using Bibliography Data. In Proceedings of the 9th International Conference on Information Visualization, pp. 953-958 (2005)
12. Inmon, W. H.: Building the Data Warehouse. Wiley Publishing, Inc. (2005)
13. Kawahara, M., Kawano, H.: An Application of Text Mining: Bibliographic Navigator Powered by Extended Association Rules. In Proceedings of the 33rd Hawaii International Conference on System Sciences, pp. 1-10 (2000)
14. Klink, S., Ley, M., Rabbidge, E., Reuther, P., Walter, B., Weber, A.: Visualising and Mining Digital Bibliographic Data. Lecture Notes in Informatics, pp. 193-197 (2004)
15. MacLennan, J., Tang, Z., Crivat, B.: Data Mining with Microsoft SQL Server 2008. Wiley Publishing, Inc. (2009)
16. MacQueen, J. B.: Some Methods for classification and Analysis of Multivariate Observations. In Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-297 (1967)
17. Microsoft Corporation: Microsoft Association Algorithm Technical Reference, http://msdn.microsoft.com/en-us/library/cc280428.aspx (2008)
18. Microsoft Corporation: Microsoft Clustering Algorithm Technical Reference, http://msdn.microsoft.com/en-us/library/cc280445.aspx (2008)
19. Muhlenbach, F., Lallich, S.: Discovering Research Communities by Clustering Bibliographical Data. IEEE/WIC/ACM International Conference on Web Intelligence, 1, pp. 500-507 (2010)
20. Mundy, J., Thornthwaite, W., Kimball, R.: The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset. John Wiley & Sons (2006)
21. Nielsen, P., White, M., Parui, U.: Microsoft® SQL Server® 2008 Bible. Wiley Publishing, Inc. (2009)
22. Olamendy, J. C.: Data mining and SQL Server 2005, http://www.c-sharpcorner.com/UploadFile/john_charles/Data_mining_and_SQL_Server_200504102006075525AM/Data_mining_and_SQL_Server_2005.aspx (2006)
23. Za¨ıane, O. R., Chen J., Goebel, R.: Mining Research Communities in Bibliographical Data. Advances in Web Mining and Web Usage Analysis, pp. 59-76 (2009)

Build a model to predict the success of students by hereditary and social factors using the tool for data mining Weka

Jasmina Nedelkoska,

Faculty of Administration and Management of Information systems, University „St.Kliment Ohridski“, Partizanska bb, 7000 Bitola, R.of Macedonia , www.famis.edu.mk

Abstract: All countries pay great attention to education and the ways of its improvement and promotion. Young people are those who will continue to build the country's future and therefore these efforts are necessary. Build a model to predict the success of students will be a great relief in the work of teachers, because they will know what to expect from a particular student. In this paper by using data from a rural environment in the Republic of Macedonia and the tool Weka two models for predicting success are built. One is using the Naпve Bayes classifier and the other using the J48 tree, a comparison between the obtained models is also made.

Keywords: Prediction, Attributes, Naive Bayes, J48 decision tree, cross-validation, folds, ConfidenceFaktor.

1 Introduction

Nowadays, lots of attention is paid in Macedonia to the modernization and improvement of education. External and internal tests have been introduced that evaluate students' and teachers' ability to work in the respective subjects. Precisely because of these reasons, we decided to work out the idea of predicting the success of students.

In our opinion, such a model for predicting the success of students might help to compare the success of students obtained in tests with the one planned earlier.

It has been proven that hereditary factors and social factors have the greatest impact on the success of the students, so we took exactly this aspect into consideration. This means that the attributes which will be used refer to these factors.

The models are constructed by using the Naпve Bayes and tree-making methods J48.

2 Description of the data set

The data used for this research are derived from a primary school in a rural environment in Macedonia. The sample is complete because all the students in the school are taken into consideration.

Since the most important factors affecting the success of students are the hereditary and social factors, we are considering precisely their influence, i.e. we examine the pattern created on the basis of these data.

Following attributes have been taken into consideration:

- Sex - male and female
- Township (where the students came from) - Brailovo, Novoselani Zapolzhani, Dolneni, Senokos, Belo Pole, Vranche, Sredorek
- Social conditions - poor, average, good
- Education of the mother - no education, primary, secondary, high
- Education of the father - no education, primary, secondary, high
- Grade - 1, 2, 3, 4, 5n, 5, 6, 7, 8 (1, 2, 3, 4 and 5n are according to the new system of education - primary education with 9 years, whereas 5, 6, 7 and 8 are according to the old program - primary education with 8 years)
- Activity (this refers to the student's activity during classes) - low, medium, high
- Intellectual development (whether the student is classified in the group of children with special needs that I marked as low intellectual development, or whether the student has some difficulty in acquiring the material with intellectual-emotional characteristics, which I marked as a medium intellectual development, finally, the largest group is students with good - normal intellectual development)
- Success (in fact the grades) - Sufficient, Good, Very.Good, Excellent

All data are of a discrete type, which means that for each attribute it's exactly known which inputs can be accepted.

122 samples have been collected, and the data are prepared in the ".arff" format so that they can be processed with the tool for data mining, Weka.

3 Description of the algorithms

Because the data is discrete, there are several methods that can be used, of these I choose the Naпve Bayes classifier and Predictive trees.

3.1. The Naпve Bayes classifier

The Naпve Bayes classifier gives us an easy access, with clear semantics, and very impressive results can be achieved. Lots of people think it is a sophisticated classifier for lots of data sets. It is the best to always start from the simplest thing. The Naпve Bayes classifier assumes that the presence (or absence) of a particular feature of the class is not related to the presence (or absence) of any other function. For example, a student is graded as excellent if he has high activity in classes, has a good intellectual development and comes from the central township

of Dolneni. Even if these characteristics depend on each other or on the existence of other features, the Naпve Bayes classifier considers all these properties which contribute to the likelihood that exactly that student is graded as excellent. The algorithm actually only counts, and then by using these numbers we assess all probabilities needed for the Bayes formula.

3.2. Predictive trees

The decision making trees are a predictive model that determines the target variable (dependent variable) of a new sample based on the different value attributes of the available data. The internal nodes of the tree indicate different attributes, the branches between the nodes show the possible values of those attributes that can be taken into account, while the end nodes show us the final value of the dependent variable.

The J48 tree classifier for decision making trees follows a simple algorithm. In order to classify a new object, a decision making tree needs to be created first, based on the values of the attributes of the available training set. So, regardless of whether it includes a set of elements (a training set) it identifies the attribute that discriminates the various samples very clearly. This feature tells a lot about data samples so that we can classify them and gain lots of information about them. But despite the possible values of this feature if there are values for which there is no ambiguity, it means that all data that are in that category have the same value for the target variable and then we need to cut that branch and to assign the same target value.

In other cases, another attribute that gives the most information can be considered. Hence, there is the need to continue in this way until a clear decision about which combination of attributes gives a target value can be obtained, or until all attributes are spent. If all attributes are spent or if a straightforward result of the available information can't be found, this branch should be awarded to the target variable with the highest probability according to the data.

Once a decision tree is obtained, the order of selected attributes prepared for the tree needs to be followed. By checking all the relevant attributes and their values with those seen in the model of the tree, we can predict the target value of the new instance.

4 Results

4.1. Naive Bayes

The cross-validation methods is selected to evaluate the chosen data, in order not to make the sets for training and testing manually. Cross-validation is a simple form that is based on significant statistical tests. Here I decided to have a fixed number of folds i.e. divisions of the data which will be used for testing and

training. The standard way to predict the error of the technique for learning is by using cross-validation with 10 folds. The data are randomly divided into 10 parts in which the class is represented with approximately the same proportion as the total samples. One part is left aside, and the scheme learns with the remaining 9, and then the error for each set is calculated. Thus the learning procedure is executed exactly 10 times, each time on different training sets. Finally, of all 10 error predictions the average error for the model is calculated.

Once the Naпve Bayes has calculated the counts that are needed for calculating the Bayes function, the following results were obtained:

For a cross-validation with 10 folds the following was obtained:

Out of 122 samples,

Correctly classified samples	73	59.8361%
Incorrectly classified samples	49	40.1639%
Kappa statistics		0.4638
Mean absolute error		0.2206
Mean square error		0.3544
Relative absolute error		58.8767%
Relative square error		81.8655%

The time to build this model is 0.02 seconds.

This model has an average precision of 0.59.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.793	0.075	0.767	0.793	0.78	0.967	Dovolen
	0.429	0.16	0.444	0.429	0.436	0.785	Dobar
	0.424	0.169	0.483	0.424	0.452	0.751	Mn.Dobar
	0.75	0.133	0.667	0.75	0.706	0.908	Odlicen
Weighted Avg.	0.598	0.135	0.59	0.598	0.593	0.851	

Fig 1. Detailed Accuracy by class made by Naпve Bayes method.

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 23  6  0  0 |  a = Dovolen
  7 12  8  1 |  b = Dobar
  0  8 14 11 |  c = Mn.Dobar
```

Fig 2. Confusion Matrix made by Naпve Bayes method.

If you change the parameters of the method - to use debugging, the assessor of the core and discretization, exactly the same accuracy is obtained, i.e. there's no change in the results.

Therefore we can conclude that with the Naпve Bayes method a precision of 0.59 can be obtained with relative absolute error of 58.87%.

3.2. J48

Once the data were released to build the tree, a tree with a size of 26 branches which have 19 end nodes or leaves was created.

The model that we build is with a cross-validation of 10 folds. The time required to build this model was 0.02 seconds. Of the 122 available samples, we obtained the following results:

Correctly classified samples 80	65.5738%
Incorrectly classified samples 42	34.4262%
Kappa statistics	0.5402
Mean absolute error	0.2319
Mean square error	0.3658
Relative absolute error	61.8969%
Relative square error	84.5021%

The average accuracy of this model is 0,641, while the relative absolute error is 61.8969%.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.828	0.054	0.828	0.828	0.828	0.872	Dovolen
	0.5	0.117	0.56	0.5	0.528	0.779	Dobar
	0.364	0.124	0.522	0.364	0.429	0.629	Mn.Dobar
	0.938	0.167	0.667	0.938	0.779	0.847	Odlicen
Weighted Avg.	0.656	0.117	0.641	0.656	0.638	0.778	

Fig 3. Detailed Accuracy by class made by J48 method.

=== Confusion Matrix ===

a	b	c	d	<-- classified as
24	4	1	0	a = Dovolen
5	14	8	1	b = Dobar
0	7	12	14	c = Mn.Dobar
0	0	2	30	d = Odlicen

Fig 4. Confusion Matrix made by J48 method.

In this model if we increase the Confidence Factor from 0.25 to 0.5, we get a larger tree with a precision less than 0,588 and with less relative absolute error of 61.6324%. If this factor is increased to 0.75 then a bigger tree is obtained, with 71 branches and 58 leaves that is built in 0.17 seconds. The average accuracy of this model is 0,612 while the relative absolute error is 56.1489%.

From this we can conclude that by increasing the confidence factor, i.e. by increasing the size of the tree, the precision of the model and the relative absolute error are reduced.

The resulting tree with a trust factor of 0.25 is as follows:

```
Intelektualna_razvienost = niska: Dovolen (21.0/3.0)
Intelektualna_razvienost = sredna
|   Angaziranost = niska: Dovolen (11.0/2.0)
|   Angaziranost = sredna: Dobar (8.0/3.0)
|   Angaziranost = visoka: Dovolen (0.0)
Intelektualna_razvienost = dobra
|   Angaziranost = niska: Dobar (13.0/4.0)
|   Angaziranost = sredna
|   |   Mesto = Brailovo
|   |   |   Pol = m: Dobar (3.0/1.0)
|   |   |   Pol = z: Mn.Dobar (3.0)
|   |   Mesto = Novoselani: Mn.Dobar (2.0/1.0)
|   |   Mesto = Zapolzani: Mn.Dobar (3.0)
|   |   Mesto = Dolneni
|   |   |   Pol = m: Mn.Dobar (2.0)
|   |   |   Pol = z: Dobar (2.0)
|   |   Mesto = Senokos
|   |   |   Obrazovanie_na_majka = osnovno: Dovolen (2.0/1.0)
|   |   |   Obrazovanie_na_majka = sredno: Mn.Dobar (4.0)
|   |   |   Obrazovanie_na_majka = visoko: Mn.Dobar (0.0)
|   |   |   Obrazovanie_na_majka = bez_obrazovanie: Mn.Dobar (0.0)
|   |   Mesto = Belo_Pole: Dobar (3.0)
|   |   Mesto = Vrance: Mn.Dobar (0.0)
|   |   Mesto = Sredorek: Odlicen (1.0)
|   Angaziranost = visoka: Odlicen (44.0/14.0)
```

Fig 5. Resulting tree with a trust factor of 0.25.

The visualized tree branches out as follows:

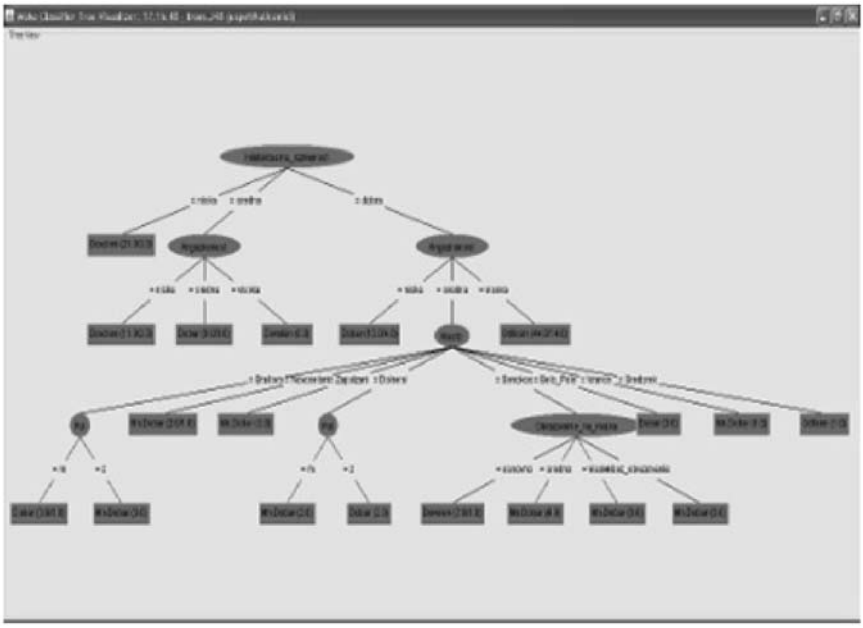


Fig 6. Visualized tree branches out.

From the resulting tree it can be seen that as the root of this tree intellectual development occurs which can be low, average and good, so therefore there are three branches.

In the low-intellectual-development group there is a very clear division, or it can immediately be seen that the success of these students is Sufficient in all such samples.

In the average-intellectual-development group there is another factor - Activity, with which the tree continues to branch out into low, medium and high activity. Here we see that the students with low activity and average intellectual development achieved Sufficient results, the students with average activity and average intellectual development achieved Good results and the students with high activity and average intellectual development achieved Sufficient results.

In the good-intellectual-development group we see a more complexly branched out tree, i.e. we see here that other factors influence success too. Students who have good intellectual development but low activity have Good results, but the pupils who have a good intellectual development and high activity have Excellent results. The students with average activity form a larger tree from which we can conclude the following: the male children from Brailovo with a medium activity and good intelectual capabilities achieved Good results, while the female children from Brailovo with a medium activity and good intellectual capabilities

achieved Very Good results. The students from Novoselani, Zapolzhani and Vranche with a medium activity and a good intellectual capabilities achieved Very Good results. The students from Sredorek with a medium activity and good intellectual capabilities achieved Excellent results, while the students from Belo Pole achieved Good results. The students from Senokos with a medium activity and good intellectual capabilities whose mothers have only primary education achieved Sufficient results, while the other students from Senokos with a medium activity and good intellectual capabilities achieved Very Good results at the end of the school year.

From the tree we can see that some attributes such as the social status, the education of the father and the grade have not been taken into account when building the tree. These factors have not been taken into account because the tree is built with a trust factor of 0.25, which means that this tree is trimmed. If a tree with a factor of 0.75 had been considered, it would have been a larger and a more complex tree.

5 Conclusion

From the results we can conclude the following:

The models using a Naпve Bayes we received:

Folds	Average precision	Relative absolute error
10.	0.59	58.8767%

From the models using the J48 we received:

Folds	The trust factor	Average precision	Relative absolute error
10.	00.25	0.641	61.8969%
10.	0.5	0.588	61.6324%
10.	0.75	0.612	56.1489%

From the tables we can conclude that the greatest precision is obtained with a model tree for prediction with a confidence factor of 0.25 and it is 0,641. The smallest error is obtained with the model tree for prediction with a confidence factor of 0.75 and it is 56.1489%.

From here we can come to the conclusion that the model obtained by J48 is better than the model obtained by Naпve Bayes.

If we see the tree we can quickly, visually infer which attributes have the greatest impact on the success of students and which attributes are left out. So we can conclude that students with low intellectual development have a Sufficient

success, the students with average intellectual development the situation is more complicated, it includes the factor activity, but we can see that the students can achieve sufficient or good results. In students with a good intellectual development the situation is more complicated because all attributes need to be considered and a branched tree is obtained.

By looking at the results we come to the conclusion that the models that we have built don't have a high degree of confidence that the error is very large. This means that from collected data a good prediction model can't be obtained.

According to me the problem is that the number of samples is relatively small - we had only 122 samples. If we take into account the attribute township - from the villages Vranche and Sredorek there are only 2 samples from Belo Pole 5 and from, Novoselani 7, I think it is an unsatisfactory number of specimens to make a good decision if in the new sample there are exactly these attributes for township. So because of this a model can't be built that will have a very small error and will be accurate.

References

1. Data Mining: Practical Machine Learning Tools and Techniques, Ian H. Witten, Eibe Frank, Mark A. Hall, Third Edition, 2011 Elsevier Inc
2. Principles of Data Mining, Max Bramer, Springer Inc
3. Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei, Third Edition, 2012 Elsevier Inc
4. http://www.d.umn.edu/~padhy005/Chapter5.html, Classification Methods

Initial Results of a Plagiarism Detection System

Atanas Semerdzhiev[1]

[1] Sofia University, Faculty of Mathematics and Informatics, Sofia, Bulgaria

asemerdzhiev@fmi.uni-sofia.bg

Abstract. During 2011 our team at the Faculty of Mathematics and Informatics of Sofia University started working on a system, which automatically tests the submitted students' solutions to programming assignments and tries to detect plagiarism. An early prototype of the system was used during the winter semester of 2011/2012 at SU-FMI and this paper summarizes the initial results.

Keywords: Assignment, Education, Homework, Plagiarism

1 Introduction

Last year our team at the Faculty of Mathematics and Informatics (FMI) of Sofia University started working on a project, which aims to develop and implement a system, which automatically tests the submitted solutions of students to programming assignments and tries to detect plagiarism [1]. The system must be able to perform tasks in the following three areas:

1. **Automatic testing of submitted solutions**. The system must be able to automatically test the solutions that our students submit to the system. Ultimately, all solutions are reviewed by the teaching staff, but the system should automate most of their work.

2. **Detection of plagiarism in the source code**. The system should be able to detect and report submissions, which share a lot of identical characteristics.

3. **Measuring solutions and detecting anomalies**. The system should be able to compute different metrics of the submitted source code and generate reports.

The system must also be able to integrate with Moodle [2], as this is the platform that we are using at our faculty.

We believe that the system will have a positive effect on the efficiency and effectiveness of the educational process at SU-FMI. It will allow us to give more practical assignments to the students, without increasing the number of working hours of the teaching staff. Please, note that the system is not intended to replace

V. Dimitrov (Editor): ISGT'2012. ISSN 1314-4855
Proceedings of the 6th International Conference on
Information Systems and Grid Technologies, Sofia, June 1-3., 2012.

the teaching staff involved in the process, but rather to aid and support their work.

During the winter semester of 2011/2012, an early prototype of the system was used in three courses at SU-FMI. The prototype was used to analyze source code written in the Scheme and C++ programming languages. This paper lists some of the results and conclusions that were obtained in the process.

2 The First Results

Although we have already used several proof-of-concept tools in the previous year, the first real use of a prototype of our system was during the winter semester of 2011/2012. We analyzed the solutions that students submitted to on-site assignments (e.g. tests) and off-site assignments (e.g. homework assignments). The system identified plagiarism in several of the submissions and all, except for one, were later proven to be real cases of plagiarism. Before the plagiarism detection system was implemented, we rarely (or not at all) observed sophisticated plagiarism attempts. As we expected, soon after it became known that we were using such a system, more sophisticated cheating attempts emerged. Two techniques stood out:

4. **Adding ballast**. In this technique one starts with two identical copies of the source code and then adds a big quantity of random source code to them. The goal is to make the ratio between identical and non-identical pieces of code in both solutions as small as possible. The random source code should of course be valid, or otherwise the submitted solution would not compile. One simple way to apply this technique is to add random (but working) functions at the beginning and/or end of the source files. We observed a more complicated attempt in which the random code was embedded in the functions of the solution themselves (i.e. the solution contained a mix of real, useful code and ballast). A very short excerpt is given in **Table 1** below. Please, note that the source code in the table is a reformatted version of the originals, as they were additionally obfuscated by making the source code formatting as chaotic as possible. The first two definitions have no meaning and in no way influence the work of the solution. Instead they are used to make the code harder to read and to try to avoid plagiarism detection. This is exemplified by the third and fourth definitions, which are essentially the same, but (1) the names of the identifiers have been changed and (2) there are calls to the reset function.

Table 1. Examples of source code with added ballast

Source Code A	Source Code B

`(define qwerty 55)`
 	
`(define (reset s)`
`  (set! s 5))`
 	
`(define (sort! L)`
`  (if (null? (cdr L))`
`  L`
`  (begin`
`  (swap! L`
`  (min-pointer L))`
`  (sort! (cdr L))`
`  L)))`
 	
`(define (sort! M)`
`  (if (null? M)`
`  (reset qwerty)`
`  (begin`
`  (sortt! (car M))`
`  (sort!(cdr M)))))`
	`(define qwertyasdfg 33)`
	
	`(define (reset s)`
	`  (set! s 3))`
	
	`(define (sortt! B)`
	`  (if (null? (cdr B))`
	`  B`
	`  (begin`
	`  (swap! B`
	`  (mmin B))`
	`  (sortt! (cdr B))`
	`  B)))`
	
	`(define (sortt! M)`
	`  (if (null?  M)`
	`  (reset qwertyasdfg)`
	`  (begin`
	`  (sort! (car M))`
	`  (sortt! (cdr M)))))`

5. Converting between expressions with equivalent semantics. This technique exploits the fact that on many occasions one algorithm can be expressed as different but equivalent (as regarding the end-results) source code constructs. You can convert between several forms and still get the same results. A very simple example is shown in **Table 2**, where an if-statement is being replaced by a cond-statement. More complex transformations may not restrict themselves to single statements only.

Table 2. Examples of expressions with equivalent semantics.

Source Code A (original)	Source Code B (transformed)
`(if (> a b) 0 1)`	`(cond ((> a b) 0)` `  (else 1))`

When the teaching staff started to analyze the submitted solutions, they not only focused on the source code, but also on the reports generated by our system and worked with them in parallel. We reached the following conclusions:

2.1 To detect and/or prove the more sophisticated cheating attempts, the teaching staff may need information, which cannot be found in one isolated submission itself.

Some examples of this type of information are given below. None of these factors alone implies plagiarism with 100% certainty. That is why they have to be used together with the reports generated from our system.

- What was the person's performance in the past? Is he/she a student, who submits a very good solution for his homework, but fails to answer even the most basic questions related to the programming language on his/her test?
- Is there a difference in the individual writing style between several submissions? For example, is there a difference in the way in which the person formats his code, picks the names of the identifiers, etc..
- Has the student made attempts to cheat in the past?

2.1 Simply giving a low score on one assignment does not prevent a student from attempting plagiarism again.

Unfortunately, it soon became clear, that if a student is caught in plagiarism, it is not enough to simply evaluate his assignment with the lowest possible score. Students who attempt plagiarism often do so systematically and some take the risk even if they were caught in the past and even if the penalty means to fail the course.

2.2 The longer and more complex the solution is, the harder it is to avoid plagiarism detection.

From our experience at FMI, we know that there are two occasions on which students tend to produce identical solutions independently from one-another:

6. When they have to write (by themselves) code that has previously been introduced by the teaching staff. For example if a lecturer in an introductory C++ course shows the students how to write a factorial function or a function that performs selection sort on an array of integers, many students will reproduce an identical copy of the lecturer's solution in their own work.

7. When the program that must be written is short. For example if you tell the students to write a simple program, which calculates factorial, there will be many independent and yet identical (or nearly identical) solutions.

We have to keep these factors in mind, when we check the submitted sources for plagiarism. On the other hand, in our practice, when we give the students an assignment, which requires them to (1) work on a problem that is new to them and (2) the resulting solution contains many lines of source code, we never receive identical solutions, except in the cases of plagiarism. I cannot give a specific number for "many lines of source code", because it varies due to many factors, including the programming language being used and the complexity of the assignment.

Another benefit of giving assignments that require the students to write more source code is that it is much harder to completely refactor a long solution. This means that on many occasions the students who attempt to cheat will only be able to process some parts of the code and leave enough pieces, which can be detected by our system.

2.1 Nothing can replace the face-to-face contact with the students

The best tool for detection of plagiarism is also the most natural one. As our experience shows, it often takes less than 5-10 minutes of conversation, to find out whether a student has really written and understands a given piece of source code or not. That is why we highly recommend that the evaluation of assignments is carried out in two phases. In the first the teaching staff checks the submissions for errors and attempts to detect plagiarism. After that a short session is scheduled, where each student has 5-10 minutes in which he is required to:

- Be able to give a short, but complete description of his solution.
- Be able to answer questions regarding the solution.
- Be able to introduce small changes to the solution.

Failing to meet these criteria, for an otherwise working solution, most often indicates plagiarism.

Performing such sessions is not always possible, but they turned out to be one of the best tools available to the teaching staff.

3 Conclusions

After running the prototype of our system, we were able to detect several plagiarism attempts, which we would have missed otherwise. As it turned out, detecting two solutions with identical characteristics is one thing, but trying to determine and prove whether this is a case of plagiarism or not is very different. We had to seek additional information and make many decisions, which cannot be made by a computer. This reinforced our belief that the system cannot be fully automated and instead should focus on supporting the work of the teaching staff.

4 References

1. Semerdzhiev, A., Trifonov, T. Nisheva, M.: Automated Tools for Assisting the Educational Process in Informatics. In: Proceedings of the International Conference on Application of Information and Communication Technology in Economy and Education (ICAICTEE) 2011, Sofia, Bulgaria, pp. 429–434. ISBN 978-954-92247-3-3 (in Bulgarian)
2. Moodle Course Management System. http://moodle.org/

An Overview of the Department of Defense Architecture Framework (DoDAF)

Vladimir Dimitrov

Faculty of Mathematics and Informatics, Sofia University, Sofia, Bulgaria

cht@fmi.uni-sofia.bg

Abstract. This paper presents an overview of the Department of Defense Architecture Framework (DoDAF). This is one of the most popular architecture frameworks and, arguably, the most popular defense industry architecture framework.

Keywords: Architecture Framework, DoDAF

Introduction

The origin of architecture frameworks can be traced back to the 80s and the work of John Zachman at IBM. At this time he was involved in the development of Business System Planning – a method which aided the management of the IT infrastructure of an organization. Zachman found that the results of his work could be applied not just to the management of automated systems, but to the enterprise as a whole. In 1987 he published his famous article "A Framework for Information Systems Architecture", which is sometimes referenced as the year of birth of enterprise architecture frameworks [1]. It took some time for Zachman to further develop his framework and it was not until 1992 that it took its well-known 6x6 matrix form. Today the framework is developed and provided by Zachman International as an industry standard [2].

Since then many enterprise frameworks have been developed. The common thing between them is that they all provide a taxonomy, which can be used to better organize information. Some of the frameworks (like Zachman's framework) provide a lot of flexibility, while others impose certain restrictions, which may be useful and guide the architect in his work. Also, some of the frameworks are created to serve the needs of a specific industry or organization (e.g. DoDAF). A comprehensive list of architecture frameworks can be found at [3].

Other important architecture frameworks in the defense industry are MODAF [4], NAF [5] and AGATE [6]. They are created to serve the specific needs of the UK's Ministry of Defense, NATO and the French government agency which makes evaluation and development programs for weapon systems for the French armed forces. As with DoDAF, these frameworks may as well be used for

civil purposes, but they are created to serve a specific industry (the armed forces) and as such some of their aspects may not be particularly useful in a civilian scenario. Their development is (to different extents) influenced by DoDAF

1 Short History of DoDAF

In 1986, a year before Zachman published his article, the US Department of Defense started working on the so called Technical Architecture Framework for Information Management (TAFIM). TAFIM was a reference model, which supported enterprise architecture development. It provided guidance for the evolution of the technical infrastructure. Meanwhile, in the beginning of the 1990s, began the development of the first version of DoDAF, under the name "C4ISR architecture framework". In 1996 the Clinger-Cohen Act was passed in the USA. It was intended to improve the way in which the government acquires, uses and disposes of information technologies. The first version of C4ISR was released in response to this act and shortly thereafter (in 1997), the second version was released.

Around this time (in 1995), the US DoD gave The Open Group permission to use the results obtained during the work on TAFIM and to create TOGAF (The Open Group Architecture Framework) [7] – another very popular architecture framework. Unlike DoDAF it does not address the specific needs of the defense industry and is adopted by organizations such as such as ING, HSBC, Intel, FedEx, UPS, etc. [8]. The current version of TOGAF is 9.1. As stated on the TOGAF site, "TOGAF 9.1 is available to all to download free of charge for non-commercial use. That will usually mean using it inside your organization." [9]. On the other hand, if you will be using the framework for commercial purposes, you have to buy a commercial license.

The decision to give the OG permission to use the results of TAFIM was (at least in part) based on two factors. First, TAFIM was funded by the government and many of its results were found to be applicable to the general public. Thus it made sense to give free access to some of the information. Second, in the late 1990s it was found that TAFIM was inconsistent with DoDAF, which was being developed at this time. As a result, TAFIM was terminated in 1999.

The development of DoDAF [10] began in the late 1990s and while C4ISR targeted the C4I community, DoDAF was intended for the entire US DoD. The first version of DoDAF was released in 2003. The current version of DoDAF is 2.02 and was released in 2010. For more information about what new features were introduced in DoDAF 2.0, see [11].

2 DoDAF Version 2

All major IT projects of the US Department of Defense (DoD) are required to develop and document architectural descriptions of the enterprise, by using

DoDAF. DoDAF serves two main purposes: (1) It provides guidance to the enterprise architects and (2) assures consistency among different architectural descriptions. This means that architectures can be measured, compared, integrated and evaluated by objective criteria and in similar manners. While DoDAF is aimed at the defense industry, it can also be used in the private, public and voluntary sectors.

The specification of DoDAF 2.0 is organized in three volumes:

- DoDAF v.2.0 Volume 1: Introduction, Overview, and Concepts, Manager's Guide [12].
- DoDAF v.2.0 Volume 2: Architectural Data and Model, Architect's Guide [13].
- DoDAF v.2.0 Volume 3: DoDAF Meta-model, Physical Exchange Specification, Developer's Guide [14].

Each volume addresses the needs of a different role involved in the creation and management of architectures. The information in a DoDAF-compliant architectural description is organized in 8 viewpoints:

1. **All Viewpoint** – Describes aspects of the architectural description, which affect all views. E.g. the context of the description, the vision, the environment, etc.
2. **Project Viewpoint** – Describes how the different programs can be grouped together to form one portfolio of programs.
3. **Capability Viewpoint** – A viewpoint, which describes capabilities on a strategic level. Intended to support strategic decision makers.
4. **Operational Viewpoint** – Describes organizations, tasks, activities and information exchanges, which are performed in order to realize the enterprise.
5. **Services Viewpoint** – Describes systems, services and their relations.
6. **Systems Viewpoint** – Describes the automated systems, their connections and system functions.
7. **Standards Viewpoint** – Describes the minimal set of rules, which regulate the work of the different elements of the architecture.
8. **Data and Information Viewpoint** – Contains conceptual, logical and physical data models. Describes the exchanged information in terms of attributes, dependencies, characteristics, etc.

An extensive overview of DoDAF can be found in Col. Mitko Stojkov's book . "Integrated system for command and control in extraordinary situations: architecture" [15].

Just like in the previous versions of the framework, DoDAF 2.0 does not enforce a specific method for architecture development. It is up to the architect to choose one. An example of developing an architectural description by using DoDAF can be found at [16] and [17]. More useful information on the practical

aspects of working with DoDAF and other architectural frameworks can be found at [18] and [19].

Acknowledgment. This paper is supported by Sofia University "St. Kliment Ohridski" SRF under Contract 134/2012.

References

1. Zachman, J. "A Framework for Information Systems Architecture". IBM Systems Journal, 1987, volume 26, issue 3. http://www.research.ibm.com/journal/50th/applications/zachman.html
2. Zachman International. http://www.zachman.com/
3. Survey of Architecture Frameworks. http://www.iso-architecture.org/ieee-1471/afs/frameworks-table.html
4. MOD Architecture Framework (MODAF). http://www.mod.uk/DefenceInternet/AboutDefence/WhatWeDo/InformationManagement/MODAF
5. NATO Architecture Homepage. http://www.nhqc3s.nato.int/HomePage.asp?session=256392830
6. Référentiel AGATE. http://www.achats.defense.gouv.fr/article33349
7. The Open Group – TOGAF. http://www.opengroup.org/togaf/
8. TOGAF Users by Market Sector. http://www3.opengroup.org/togaf/users-by-market-sector
9. TOGAF Version 9.1 - Download. (last acces 14.04.2012). http://www.opengroup.org/architecture/togaf91/downloads.htm
10. DoDAF Architecture Framework Version 2.0. http://dodcio.defense.gov/dodaf20.aspx
11. Semerdjiev, A. Department of Defense Architecture Framework (DoDAF) ver. 2. In the proceedings of the Annual Scientific Conference of the Vasil Levski National Military University. Veliko Tarnovo: 2010, pp. 23 – 30. ISSN 1314-1937 (in Bulgarian)
12. DoDAF v.2.0 Volume 1: Introduction, Overview, and Concepts, Manager's Guide. 28 May 2009. http://cio nii.defense.gov/sites/dodaf20/archives.html
13. DoDAF v.2.0 Volume 2: Architectural Data and Model, Architect's Guide. 28 May 2009. http://cio nii.defense.gov/sites/dodaf20/archives.html
14. DoDAF v.2.0 Volume 3: DoDAF Meta-model, Physical Exchange Specification, Developer's Guide. 28 May 2009. http://cio nii.defense.gov/sites/dodaf20/archives.html
15. Stoykov, M. "Integrated system for command and control in extraordinary situations: architecture". Softtrade, Sofia, 2006. ISBN 9789543340392 (in Bulgarian)
16. Semerdjiev, A. Designing Organizations – Development of a Functional model. Sofia: "St. Georgi Pobedonosec" Military Publishing House, 2006, Military Journal, Vol. 5, pp. 145-158. In Bulgarian. ISSN 0861-7392 (in Bulgarian)
17. Semerdjiev, A. Designing Organizations – Development of a System model. Sofia: "St. Georgi Pobedonosec" Military Publishing House, 2006, Military Journal, Vol. 6, pp. 105-119. In Bulgarian. ISSN 0861-7392 (in Bulgarian)
18. Dimov, A., G. Stankov, and T. Tagarev. "Using Architectural Models to Identify Opportunities for Improvement of Acquisition Management." Information & Security : An International Journal 23, no. 2 (2009): 188-203.
19. Napoli J.P. K. Kaloyanova. An Integrated Approach for RUP, EA, SOA and BPM Implementation, Proceedings of the 12th International Conference on Computer Systems and Technologies, Vienna, Austria, 16-17 June 2011, ACM DL, ISBN: 978-1-4503-0917-2, pp.63-68

INTELLIGENT SYSTEMS

Digital Spaces in "Popular" Culture Ontologized

Míchéal Mac an Airchinnigh

School of Computer Science and Statistics, University of Dublin, Trinity College Dublin, Ireland
mmacanai@cs.tcd.ie

Abstract. Electronic publication seems to have the biased connotation of a one-way delivery system from the source to the masses. It originates within the "Push culture." Specifically, it triggers the notion of (a small number of) producers and the (masses of) consumers. The World Wide Web (but not the Internet) tends to reinforce this intrinsic paradigm, one that is not very different from schooling. Wikipedia, in all its linguistic varieties is a classical publication of the collective. Much enthusiasm is expressed for the English version. The rest of the other world languages do not fare so well. It is of interest to consider the nature and status of Wikipedia with respect to two specific fields, Mathematics (which is not so popular) and Soap Operas (popular for the masses, largely sponsored by the Corporations).

Keywords: Digital Space, Intelligent Artificial Life, Ontology, Soap Opera

1. The Talking Drum

"Before long, there were people for whom the path of communications technology had leapt directly from the talking drum to the mobile phone, skipping over the intermediate stages" [1][2]

It comes as a shock for most, to know that the first successful transmission of message data (i.e., significant amount of meaningful information, in the technical scientific sense) at a distance, was done on African drums? From the book cited above one may, with clear conscience, jump to Wikipedia to see if there is information on the "Talking drum" [1, 3].

The previous sentence structure is carefully crafted for those who believe that the printed book has priority/superiority over a distributed elusive-multi-author, multi-linguistic electronic resource. On the other hand we will look at the potential problem of the multitudinous electronic resources that offer a cornucopia of Information in most natural languages of the entire world. To explore this hypothesis we will focus on two cultural items: a) Mathematics, a topic for the few, significant for all, and b) Soap Operas, a topic for the many, interesting for most.

What is known? What can be known in the Digital Culture of our times? The visual seems to dominate and appears more accessible to us in 2012 than either the word or the speech? By visual one means, the film, the TV programme, the picture,

the poster, in other words that thing towards which the eyes are immediately drawn. In certain cultures and at certain times, certain kinds of images have been strictly forbidden [4].

Like it or not, Wikipedia, is an enormous resource for both Mathematics and Soap Operas. It is frequently the case that a Wikipedia topic is presented inmore than one natural language. In general, a topic in the English Wikipedia has correspondents in other language versions. On the other hand, there are topics in non-English Wikipedias that have no correspondents in English. Consequently, it is of importance to note that Google Translate can give a reasonably good sense of the content of a Wikipedia page written in language L (with translation into M).

Now is a good time to introduce our naming convention for different language versions of Wikipedia. For the big English Wikipedia one might obviously write WP. But there are many language versions that would have the same abbreviation. Hence one further suggests the use of the international forms "en" and so on to distinguish between distinct language versions. For example, to denote the English and Portuguese versions one might write enWP and ptWP, respectively. For Cyrillic, and especially for the Bulgarian Wikipedia, one proposes the use of bgУП, preserving the current English prefix, but respecting the Cyrillic abbreviation.

Later on one will have a similar naming problem to resolve for actors and the roles that they play in particular Soap Operas. Again, the natural language versions need to be multi-faceted, if only to be harmonized with respect to the usual sorts of accounts given in the various linguistic versions of Wikipedia.

Similarly, is one not also aware of the apparent dominance of music within the general realm of sound. Might there not be a good reason for an aural uprising against the "digital sound"? [5]. On a speculative note one recalls the significance of the digital musical sounds for communication in "Close encounters of the Third Kind" [6].

Wikipedia has a very high profile in the English-speaking world. Access to a particular page in enWP ([en] may be omitted on the understanding that, by default, one is accessing the English version of WP), on a specific date, may be registered in the standard ISO format [yyyy-mm-dd]. For example, let us now turn to enWP[2012-01-09] to see what kinds of interpretations are currently given to "digital shape". Naturally, one will compare search results with the more primitive "shape".

There are Wikipedia competitors of all sorts in 2012. For example, a search on the phrase "Criticism of Wikipedia" will bring up many interesting WebSites, of which Knol [7] is a special one, now terminated [8].

The phrase "digital shapes" conjures up, in the first instance perhaps, geometrical digital shapes? In other words, in this context the digital shape in

not amorphous. It has a certain familiarity. If we imagine, a triangle, a square, a rectangle, a circle, we conjure up in our minds the corresponding familiar shape. To speak of the Pentagon, with deliberate capitalization, recalls the shape of that building of the same name in Arlington County, Virginia, near Washington D.C., and separated from same by the Potomac river. Such specifically geometrical shapes lie at the very foundation of all civilizations and the artefacts that arose therefrom.

2. Mathematics

Whether one likes it or not, Mathematics is a core cultural phenomenon in humanity. To ontologize Mathematical material is, in a deep sense, easy! I mean to say, to mathematize is, de facto, to ontologize reality. Let me tell you about the Masters Course that I teach?

Students taking the MSc course in Interactive Entertainment Technology (MSc IET), in Trinity College Dublin, are currently required, not only to use Wikipedia to ascertain relevant mathematical information, but also to register formally as editors of Wikipedia, the better to understand something of the process of maintaining a specific body of digital knowledge. In addition to Wikipedia, the students are also required to use Wolfram Alpha. Scholarpedia [9, 10] is on a different "higher academic" level. It is not considered to be a primary electronic mathematical resource for the MSc IET course. It does have its uses and students are recommended to use it for certain topics.

Examination of the subject is "open book." The students use Mathematica and Wolfram Alpha for all their practical work. The examination paper is currently delivered in classical paper format, in order to comply with the University of Dublin Regulations. In practice, the paper is an electronic publication (typeset with LaTeX [11]) and printed on paper to conform with current regulations. The solutions to the examination paper are developed in electronic form (specifically as Mathematica notebooks). Such solutions are reprocessed in pdf form for electronic transmission to the external examiner (currently elsewhere in the EU). The students answer the questions with the use of the Mathematica notebook form. The notebook is then e-mailed to the examiner, and in addition, copied to a USB stick as backup.

en (3,787,457) (3,840,318)	bg (122 840) (125 409)	pt (703 219) (709 567)	fr (1 168 548) (1 195 043)	de (1.307.907) (1,339,416)
Square (geometry)	**Квадрат**	**Quadrado**	**Carr ìe**	Quadrat (Geometrie)

Laplace transform	Трансформация на Лаплас	Transformada de Laplace	Transform ìee de Laplace	Laplace transform
z-transform	**Z-преобразование**	**Transformada Z**	Transform ìee en Z	**Z-Transformation**
Cellular automaton	ru Клеточныи автомат		Automate cellulaire	Zellul arer Automat
Spline (mathematics)	**ru Сплаин**	**Spline**	**Spline**	**Spline**

Table 1. Some MSc IET mathematical topics of the first semester

Consider some of the specific mathematical topics taught in the first semester (September-December 2011): Laplace Transform, z-transform, cellular automaton, and splines. The students often have a second or third language. Hence, for each topic they are encouraged to look up another language equivalent. A very brief summary of the sort of work done is illustrated in Table 1. The header row gives the language. The first row illustrates a Wikipedia lookup for a simple mathematical concept, the square. The subsequent rows illustrate four of the key mathematical concepts used in the first semester.

NOTES on MSc IET mathematical topics of Table 1:

1. In the original research, one looked at the Irish language website, ga (13,478), with a view to ascertaining its suitability for "Higher Level Mathematics." The results were so appalling that the data has been deliberately omitted. One may conclude, that in the matter of Science and Mathematics, the Irish language culture is apparently in significant decline, from the point of view of the Digital Space. The interested reader will note that, with respect to Wikipedia, many world cultures are in a similar position.
2. In the second column one started with the Bulgarian Wikipedia (bg); the mathematics pages were too few, inadequate; one switches to the Russian Wikipedia (ru) for better comparison. There are roughly 7 times more ru pages than bg.
3. In the last column one started with the Irish Wikipedia (ga); the mathematical entries were practically non-existent. One might use the nature and amount of Mathematical pages in Wikipedia to judge the health of the corresponding culture.

Mathematics is one of the keystones of any culture. It is the foundation of any science, whatsoever. Not only is it pragmatic and aesthetic, but also practical and publishable. The nature of the (electronic) publication of Mathematics in our times is due to one man, Donald Knuth [12]. He it is who gave us TeX and MetaFont. In practice, today, one uses LaTeX [11] or, considering all the other

languages on earth, the XeTeX of Jonathan Kew. Of particular importance, for the publication of Mathematics within Wikipedia, is the use by MediaWiki of a subset of AMS-LaTeX markup.

From the digital space point of view, one also needs to mention the use of online video clips of lectures, whether delivered via YouTube or on specific College sites such as MIT Opencourseware [13].

3. Digital Space/Digital Shape

Now let us turn to what might seem to be an easy application of Wikipedia, with respect to these concept words?

E75_Conceptual_Object_Appelation: Digital_Space
E75_Conceptual_Object_Appelation: Digital_Shape

en	bg	pt	fr	de
Shape Start-Class				
Space C-Class	Пространство	Espaзo fнsico	Espace (notion)	Raum (Physik)
Digital culture		Era da Informaзго	Sociйtй de la connaissance	Informationszeitalter

4. Digital story telling: Забранена любов (Forbidden love)

Let us consider the Soap Opera [14], (and related Telenovela [15, 16], a pulp-television show) in its digital form, abbreviated here as Digital Soap? It is a particular modern form of story telling, originally born in the age of radio (product placement of soap by Procter and Gamble [16]) and developed extensively in the age of television. In our times, the latter has migrated into the digital space. From the point of view of the Digital Space it is of considerable significance that one can still experience something of that original Soap Opera today. The soap product was OXYDOL [17] and one can still listen to the original broadcasts [18].

For definiteness we shall use a Bulgarian Soap, entitled "Забранена любов (Forbidden love)", which is accessible on VBOX7 and formerly on YouTube. The Soap is originally based on an Australian Soap entitled "Sons and Daughters" [19]. We will need to distinguish between the actor and the role played. We will need to manage the names of the actor and role in both Bulgarian and English. We will want to show pictures of the actor and of the role played for this specific Soap. Such pictures might cause certain difficulties with respect to copyright. From a research point there is always "fair use". One way in which to manage one's research is to set up a private Flickr group where the picture may be viewed but

cannot be reused. There is such a private Flickr group with the name "Забранена любов" [20]. Naturally, one will want to tag the pictures with appropriate information. A more formal approach can be taken with use of an ontology editor such as Protégé [21]. See, for example, "Keyimages ontologized for the Cultural Masses" [22].

A key question for us is to determine the boundary of the digital space of Забранена любов. For example, in addition to the above, there is a Croation version, "*Zabranjena ljubav"* [23]. A second key question is to determine to what extent it corresponds to similar Soaps currently available in the Island of Ireland and correspondingly to consider to what extent the latter might be "translated" to countries such as Bulgaria, Turkey, Portugal.

Soap	TV Host	Wikipedia	Star XX	Star XY
Забранена любов Forbidden love	NOVA	bg, en	Ева Захариева Eva Zaharieva	Борис Костантинов Boris Konstantinov
EastEnders	BBC	cy, de, en, eu, fr, ga[x]		
Fair City	RTÉ	en, ga, pl	StarXX	StarXY
Neighbours	FremantleMedia	en+15	StarXX	StarXY
Ros na Run	TG4	en, cy, ga, eu	StarXX	StarXY
Emmerdale				
Coronation Street				
Cougar Town				
Doctors		—		
Стъклен дом Glass house	BTV	bg, en, hr	Елена Атанасова	Христо Атанасов

Table: NOTES on the Soap Operas

1. Star XX and Star XY [24] denote key female and male actors/roles; for example Ева Захариева (Яна Маринова) и Борис Костантинов (Тодор Танчев). In particular, note that for passionate viewers of the Soap, the Soap person identity is the "real name" by which (s)he is known to the public. If the Soap runs long enough the actor's real name becomes irrelevant. In other words the actor and the part become one, the Soap name becoming the "real name" of the actor. This is a major "event" for the Ontology!
2. There does not appear to be a specific website for Забранена любов any longer. One deduces that a significant part of the Soap's Digital Space has been excised.

3. It is worth noting that there is a similar "corresponding soap," entitled Забраненият плод [25] (Forbidden fruit, originally Turkish (Istanbul setting), and dubbed for Bulgaria.)
4. Other entries are deliberately blanked. The reader is invited to fill in appropriate details according to taste. In other words, the Semantic Field of Google + Wikipedia + WhateverOneWants is sufficient to find out most relevant information in most languages of the World, today!
5. The last entry in the table records another successful (and more polished) Bulgarian Soap opera. Its inclusion in the table is intended to provide for a little bit of fun for the reader. Specifically, one is invited to make the connection between "Forbidden love" and "Glass house."
6. Other entries are deliberately blanked. The reader is invited to fill in appropriate details according to taste.

4. Everyman as publisher

Schooling is one way in which to force a person (qua student) to confront the possibility of making public their thoughts (originally orally). Such initial self-expression has many possible outlets: teaching, acting, and so on. After the oral there comes the writing, the author. Now in our Digital Age, there are multi-media in which one can express oneself, to show and tell. The audience for such media is virtual (and maybe non-existent). It is the story which is the thing. That story may unfold orally, in writing, in clay, in paint, or in our times, digitally. The mathematical shape of the parabola, may manifest itself as a sketch or as stone bridge. What might be the electronic equivalent? What sort of tools does one need?

The research goal for 2012 now becomes very specific. It is to assemble a collection of digital devices (such as phone and tablet), a collection of digital spaces (such as music, video, and still image), a collection of knowledge sources (Wikipedia, Books, Articles) and then to construct and demonstrate a new kind of Dynamic, a Digital Shape in the field of Electronic Publication. It is nothing more than the manifestation of a story (in the classical sense), determined by Digital Means. In particular, this research intends to reverse the action, from passive story feed via TV to active **deconstruction** via the person and her/his digital artefacts. There is a phrase for such activity, remix culture [26]. In his book~RemixEconomy Lawrence Lessig cites the now classic case of a mother posting a YouTube video clip (29 seconds) of her baby boy dancing to some background music, of someone called "Prince".

Let us begin and end with the Soap opera? Let us tear it apart? The key to deconstruction is to capture stills from the dynamic video. This requires bookmarking or tagging of video frames. Given a character, can one re-identify

the same character in a different setting? Is this sequence just like Romeo and Juliet? Perhaps that one is pure Macbeth? How can I show and tell you what I have seen, discovered, deconstructed? How can I ePublish it? For whom?

5. Conclusions and Suggestions

If a Soap Opera runs on, and on, and on, then the magic of the story tends to disappear and become banal. It takes on the characteristics of daily life, albeit in another form. The Electronic Publications of 2012, and beyond, must be of the same type, short enough to be interesting, to tell a story. Wikipedia captures much of the nature of the kernel this kind of story. There is text, there are images, there might be some video links. There is background explanation of how the story came to be. This paper suggests how one might take this one step further, with the use of a collection of personal electronic devices to remix and publish at will.

6. Acknowledgements

The recently published book, "From Gutenberg to Zuckerberg, What you really need to know about the Internet" by John Naughton 2012, arrived in time to help me re-think the course of Ontological History! Of particular interest was his short section on Wikipedia, p.88-96, which confirms the nature of Electronic Publication, in general, and happily correlates with the thesis formulated in this paper: the reader as activist writer, correlator, publisher, in short, Everyman.

References

1. Gleick, J., *The information : a history, a theory, a flood*. 1st ed2011, New York: Pantheon Books. 526 p.
2. 1 | DRUMS THAT TALK (When a Code is Not a Code) p.27
3. Wikipedia Editors *Talking Drum*. 2012.
4. Wikipedia Editors, *Iconoclasm*, 2012.
5. Wikipedia Editors. *Digital audio*. 2012 [cited 2012 2012-01-11]; Available from: http://en.wikipedia.org/wiki/Digital_audio.
6. Wikipedia Editors. *Close Encounters of the Third Kind*. 2012 [cited 2012; Available from: http://en.wikipedia.org/wiki/Close_Encounters_of_the_Third_Kind.
7. Google. *Knol, A unit of knowledge*. 2012 [cited 2012 2012-01-11]; Available from: http://knol.google.com/k/criticism-of-wikipedia.
8. Google. *Knol termination*. 2012 [cited 2012 2012-01-11]; Available from: https://knol-redirects.appspot.com/faq.html.
9. Wikipedia Editors *Scholarpedia*. 2012. **2012**.
10. Expert scholars. *Scholarpedia*. 2012 [cited 2012 2012-05-19]; Available from: http://www.scholarpedia.org/article/Main_Page.
11. Wikipedia Editors *LaTeX*. 2012.

12. Wikipedia Editors *Donald Knuth*. 2012.
13. MIT. *Open Courseware*. 2012 [cited 2012 2012-05-19]; Available from: http://ocw.mit.edu/courses/mathematics/ - grad.
14. Wikipedia Editors *Soap Opera*. 2012.
15. Wikipedia Editors *Telenovela*. 2012.
16. Уикипедия *Теленовела*. 2012.
17. *OXYDOL'S OWN MA PERKINS*. 1933.
18. Frank & Anne Hummert *Ma Perkins 01*. Internet Archive, 1933.
19. Wkipedia Editors *Sons and Daughters (Australian TV series)*. 2012.
20. Михал Орела. *Забранена любов*. 2012 [cited 2012 2012-05-20]; Available from: http://www.flickr.com/groups/zabranena_lubov/.
21. Stanford Center for Biomedical Informatics Research. *Protŭgŭ*. 2012 [cited 2012 2012-05-20]; Available from: http://protege.stanford.edu/.
22. Mac an Airchinnigh, M., *Keyimages ontologized for the Cultural Masses.* Information Systems and Grid Technologies, 2011: p. 95-106.
23. Wikipedia Editors *Zabranjena ljubav*. 2012.
24. Wikipedia Editors *XY sex-determination system*. 2012.
25. Wikipedia Editors *Забраненият плод (сериал)*. 2012.
26. Wikipedia Editors *Remix*. 2012.

Application of knowledge management information systems in digital redactions

Elena Miceska

Faculty of Administration and Information Systems Management

University "St. Kliment Ohridski"- Bitola, Republic of Macedonia

Abstract. Information is the lifeblood of every redaction. Basically the information is the raw material for the reporter, but on a higher level of newsroom management, editors require more frequent knowledge about how to manage the information they receive. Nowadays with the information overload newsroom organization are forced to reorganize to face with the information in another way. In this paper will be explained remaining factors that are the key drivers for change in journalism in the 21st century to satisfy his customers. In this information age has several tools and software that is available to journalists. At the end it will be discussed the idea of applying the Intranet as a mean for knowledge sharing within the redaction.

1 Knowledge management in journalism organizations

What is called knowledge management? Most scientists the knowledge management have been defined as the art of creating value by extending the organizational intangible assets. Everything that journalists collect, share and publish represents most valuable asset for journalistic organizations. The ability of journalistic organizations to distribute stories that are unique and are very difficult to imitate reflect their competitive ability. This capability is the simplest form of knowledge management within journalism. Journalists create knowledge in the process of synthesizing information and expression of quality and integrity of the stories. Ifra (www.ifra.com) is a collection of over 1,800 journalists and tech companies that are engaged in research on the impact of technology on journalism. According to Ifra, the transformation of journalistic companies in knowledge management companies depends by how its editors and journalists are willing to exploit and manage the knowledge they have. Successful news organizations need to use knowledge management as an intangible asset to improve their journalism which results in more tangible benefits.

Information overload. Media companies have been successful if they have the opportunity to worship the information on their customers. Information required by her readers and listeners to work on it, or to extract something from it that is useful for them. Some of the roles of journalism are to turn information into knowledge

(to make it more useful for its audience and find a way to preserve knowledge for future use). Therefore one of the reasons for change in journalism is the surge of information. The problem consists in the abundance of data undermines the attempt to quickly make important decisions. Essential for redaction is to understand the radical difference from lack to excess of information, because this problem is much more serious than the pursuit of technological change. The emergence of excess information is a result of information today that much easier to produce, but the human capacity to absorb them slowly changing. Powerful technologies such as satellites, 24 - hour television and the Web are the most powerful information sources and causes of excess information. Today, it shows that people using the Internet can come very quickly to the news as opposed to journalists.

The new technology a major reason for changes in journalism. Technology is the most important driver of changes in journalism over the past two decades that will continue to transform journalism. Most journalists consider that journalism today is different because of the impact of technology. The biggest proof of this is the way of access to information, the processes used for their processing and devices used for their preservation. It is necessary to emphasize that technology should not control the selection. Instead, it provides many options for many people who are responsible for making important decisions and thus should be used as a tool to improve journalism today. Though journalists are considered as part of elite for knowledge management, however their membership primarily depends on their proficiency with technology, because the technology is one of the reasons for the existence and access to large amounts of data over the Web, satellites, cellular phones, electronic mail and other information and communication technologies.

The Internet as a major catalyst for change. The Internet is considered as a major catalyst for change. Essentially, the Internet is the catalyst of the historical transition from one age to another. A collection of information and its transformation into knowledge is much more important in this information age than the previous industrial age. The Internet is the main culprit that makes people expect free information. Therefore bad news for journalists and journalism is that journalism organizations mutually can compete only in terms of their reporting (how it may be better), but it requires large deposits and investments. However, there is a good news and that is that journalism and journalists may become more important than ever before in this information age.

2 Redefinition of the redactions

To be able to develop knowledge management within a redaction should be made three major changes. The first and most important change is the change of

mindset of journalists and how journalists see their work. The news should move from a certain specific form to a variety of forms for different customer demands due to changing their lifestyle. Journalists should be prepared to change the way they work that involves a new view of the role they play. The second change is related to the physical structure of the redaction. The geographical position directly affects the flow of information, so redactions should be restructured in order to facilitate the flow of information. The third change relates to technology used by journalists and their attitude towards it. Journalists need to embrace the benefits of technology. These three changes require a change in the way of functioning and organization of the redaction.

Changing the mindset of redaction. Changing the mindset of redaction means changing individual considerations and corporate culture of redaction. Changing the mindset is related to several factors including different attitude in terms of time, a flexible approach to journalistic role and readiness to adopt more collaborative forms of work. Knowledge management is necessary to implement a new type of redaction that is required for multiple reporting. It means "overhaul" the editorial position of the information that it handles and knowledge development. To be able to make newspapers more skilled and responsible, redactions should be developed as real multimedia companies and organizational structures. The purpose of changing the mindset of the redaction is to all individual minds in the company to function as a collective journalistic mind.

Change of the physical structure of the redaction. Physical reorganization of redactions is needed for an impact on the work of journalists. Newsrooms around the world make change in their physical organization because traditional structure and organization of the redactions inhibits creative thinking.

Newsplex is an example of how it should look like a redaction in the new era of knowledge. Media architect Saf Fahim, has designed Newsplex as a model by which journalistic publishers can learn. In the drafts of Newsplex, Saf Fahim took into account the intelligent building, intelligent systems and intelligent cabinets to develop a work environment that will function as a real and virtual site for multimedia experimentation. With that Newsplex enables more productive and adaptable work environment for modern journalists. On the lowest floor of the construction of Newsplex is the area for the most current news including the central office for information flow. The above are rooms for individual journalistic research and reporting. Within such a work environment convergence journalists have access to the latest and most useful modern information tools for information management. The function of the main desk is to coordinate news and functions for editorial management knowledge. The line of specially designed and highly flexible jobs supports multi-media reporting and production activities of the news. As a further improvement of Newsplex is wireless networking at very high speeds. Modern communications packages that include mobile and video conferencing

allow journalists to be in constant contact with each other and access to digital infrastructure. On top of this infrastructure is set database server that has the ability to accept, to categorize and distribute all types of formats of journalistic materials. Newsplex infrastructure is designed to be able to provide easy, simple and effective way of connecting the new technological tools to be tested and used. All equipment and facilities available to Newsplex are movable and configured to be able to provide dynamic reorganization of the work environment. The purpose of this organization of Newsplex is to relieve of the walls that divide people. Within the work environment is set up a screen that displays the status of readiness of each page of the newspaper or activity. In this way each employee can be up to date with developments and to propose any suggestions in order certain activity can be successfully completed.

Finally we can conclude that the architecture of the editorial board that is open, friendly set and which allows free flow of information, monitor journalists' needs for new technologies and tools and timely suits them, presents a real source of ideas.

Implementation of new technology. In the last three decades, most newsrooms have embraced the digital evolution. Digital cameras and the ability to scan enable development of photo-journalism. Satellite technology and digital connections allow newsrooms to receive huge amounts of data very quickly. The emergence of the Internet further increases the flood of information available to journalists and nowadays it is one of the problems with which they deal, the flow of information and lack of space for their preservation. For one edition to be called digital is not only required to implement new technology, instead there is counted the organizational culture, a way of defining critical processes that face with journalistic operations and a way of linking these processes which makes redaction dynamic. With the implementation of new technology and its efficient utilization in the future number of employees within a redaction would be substantially decreased. Additional redactions work on the development of large knowledge bases based on the possible greater amounts of information and implementation of virtual redactions from which data will be available to all employees from anywhere in the world. Consequently, each new technological device tends to become a catalyst for defining a new type of journalism.

3 New tools for sophisticated journalists

Technological development allows a number of tools used in journalism in order to improve it. These include refined management information such as computerized reporting, geographic information systems and research related to how databases are changing the way of functioning of new journalistic organizations. The key to better journalism in the new information age is appropriate

and continuous training and providing new tools to reporters in combination with dedicated leadership. In other words, the best way that journalistic organizations can cope with technology and the flow of information is effective utilization of technology.

Successful managers of data. Since previous studies which have been made, the surge of information that appears in the new information age, does not bother the people who are familiar with the technology. Anyone who knows effectively use electronic mail and how to move through the data can deal with flood of news and information, are characterized as info-coper, a person who is satisfied with the technology and information overload. Therefore it is necessary, all journalists be trained to be able to recognize the benefits and disadvantages of using e-mail packages.

NewsGear - technological advanced journalism project. NewsGear (www.newsgear.info) is project initialized by Ifra to investigate the tools available to journalists so that they can operate in multi-media and fully digital environment. With him each year are evaluated hundreds of products and technologies. He really is a mobile redaction that reporters provides tools and opportunities they need to cooperate with their colleagues, managers and access to resources they need while creating each new story.

Web Reporter is a new tool which is developed by a young German journalist, Ellen Tulickas, which for the first time created a journalist report, complete with pictures directly on the website. His equipment consists of a headset with a small monitor in the middle of the left eye and a digital camera in the middle of the right eye, a miniature computer and battery with a belt which is fastened around the waist. The report was written by keyboard slung around her wrist. This invention allows full mobilization of the journalist, availability in real time and direct reporting from the scene.

Today there is a new trend in journalistic organizations in which journalists are supplied with portable equipment (laptops, digital cameras, camcorders, mobile phones) for getting the news that enables greater mobility and independence of the redaction. If the device has an easily portable size, can be used as a powerful device for collecting information. It can provide fast content reporting, resulting to a greater accuracy. With the assistance of broadband access and wireless application protocol can accelerate the ability of the device for generating information.

Computer-based reporting. If unique and relevant content is the condition under which the determination is made about the quality of a journalistic organization, then journalistic organizations should invest in ways that will enable generating unique and relevant content. One of the ways in which can be generated exclusive content is the use of geographical information systems.

Geographic information systems (GIS) as a combination of cartography and

digital databases that work together under the cover of today's desktop computers, produced maps and associated statistics that show reporters where is the place where the event occurs, perform the ranking of criminal incidents, display demographic data for illustration of flooding or pollution of the environment and the like.

Modern journalists can use the maps that generate GIS in order to understand various phenomena and make a story about them. GIS allows journalists additional explanatory elements that can add to their statements, as mapping software provides a new way to utilize information. In the future expect a growing number of journalists to attend training on the use of geographical information systems where they will perceive the benefits of using GIS.

Databases for information management. With the help of software for databases, newsrooms can create useful and search databases by the mass of information they gather daily. Relational databases can be used to store the information posted while flat file databases can be used for information not published, as contact with a source of news and research notes. By effectively integrating the two types of databases into a single system, newsrooms would have a powerful database that will support printing and electronic publishing, including managing the process of issuance. Databases and management information functions are necessary to store and use all information that are available in the redaction. Moreover they need intelligent systems, so that when journalists will have a task that would be similar to a story which have already created and saved in the database, the system will pull the information for that story and it will show the journalist. This means that newsrooms should use bases of knowledge.

Convergent journalism. Because today's redactions use a digital picture receiver and use systems that can process the multi-medial web documents, we can say that they are a part of convergent journalism. XML is a language which allows the construction of convergent journalism. It describes how data can be displayed in many formats of data displayed on PDA screens or mobile phones to data printed in newspapers or stores. XML defines a standard for the creation of standards for exchange of information. The processor of XML Stylesheet Language (XSL) is a combination of an XML document and data interface or XSL rules for transforming and reformatting the data for adapting the device through which should be presented. XSL provides an automatic transformation of documents to another form. For example, if a story written for a particular page of a newspaper, can be reformatted to be displayed on a private device for data presentation. This means that XML allows to reporters to write articles once and to announce anywhere. In addition, XML provides a standard way to exchange content that arises and is designed for multiple media.

Five essential tools for enabling mobile journalism. Today all that need

to make a journalist to publish his story is to make just a few clicks due to the available tools that allow full mobility and independence of the redaction.

Voice recorder/Google voice (itunes.apple.com), constitutes a useful alternative of telephone conferences. The subscribers of Google voice are given a unique number to divert calls from mobile or landline phone with the ability to store voice messages on an easily accessible website. The best feature of this recorder is the ability to capture integrated speech for incoming calls (like voice mail) and the ability to access the conference from anywhere via the Internet.

Ustream (www.ustream.tv) is an iPhone application for live streaming (tracking). Users of the popular video site Ustream can pose and send audio and video content in real time. In addition there is the possibility of using the Social Stream to integrate experiences in popular social websites. Depending on connection, Ustream Broadcaster can be quite variable.

Reeldirector (www.nexvio.com) is a package designed for the iPhone that allows video processing. On Reeldirector can be attached video clips, to reduce the size of the videos, to add a text, to incorporate sound and the like. Although journalists would often like to use this package for editing and broadcast their video content, however this application is suitable for mounting quick and short video interviews or key events.

Fast Thumbs & Sync-able Memos (evernote.com). This application serves as a substitute for reporter's notebook and pen to write some notes during their research on their iPhone. With this tool further increases the reliability of the written story. By synchronizing the written story of a server by using applications such as Notespark and Evernote, if will be lost the device on which is placed the story she will still be a protected and available to journalists on the server that is stored.

WordPress (wordpress.org) is an application designed for the iPhone that allows reporters to their analysis and multimedia content to pack together without using a computer. Even if certain journalists should consult with the editor for publishing news, WordPress can give the editor a working schedule for how to follow the news.

4 Intranet as a tool for distributing knowledge management

One of the key ways of applying knowledge within the redaction is through teamwork and cooperation. An intranet is a very powerful tool that enables teamwork, cooperation and exchange of ideas. With proper use and effective maintenance, intranet allows employees to better communicate, explore and manage their knowledge.

When it comes to knowledge management is believed that information becomes very valuable when they are shared. Journalism is a profession that

usually ignores cooperation. But this culture of newsrooms must change in order to encourage the cooperation of journalists. As part of their culture, newsrooms tend to save as much as possible the information they receive. The journalists must know what type of information and in what form they need so they can take to achieve their goals. If the intranet is connected with the research databases access to historical or archived data is further facilitated. However in order to function Intranet within an redaction, editorial managers need to implement the model of knowledge sharing within the organization and allow free exchange of information.

Also, an intranet is a very useful tool for keeping current data and documents that need frequent changing. It is much cheaper mean of distribution of documents, as opposed to their distribution in print or postal paper version. Because Intranet uses the same network infrastructure and protocol as the Internet, many communication tools as faxes, email and remote access to remote computers are available for the journalists at very low price.

Another of the benefits of an intranet for journalistic organizations are saving time with him. If there is shock news that editors should know, they can immediately be published. Some redactions set their intranet sites to be updated every hour. Advanced news organizations create virtual redactions. Every journalist has a password with which allows access to the virtual editorial board or to the main web page of editorial. After logging in, the journalist has many benefits: can see his tasks for the day, the list of his duties, SMS and e-mail.

5 Conclusion

Because content that offer journalistic organizations is their greatest asset, then they should cherish, and it requires new skills and approaches to the new era of knowledge. For this purpose it is necessary to develop knowledge management in each redaction. It requires changes in thinking on the redaction and the way journalists see their profession. In order to develop knowledge management in the redactions, it is necessary to change the journalist's attitude towards technology and continually accepting and getting used to new tools for the job. Convergence journalism can be considered as the initial phase of implementation of knowledge management. It deals with redefining the redaction in order to find a way that editors will deal with the flood of information. Redactions use the Intranet as a mean through which they share their knowledge and a way in which they connect their technologies to generate better journalism. Today, in the new information age has many tools that are available to journalists, which facilitates their work and increases the quality of journalism. Some of them are computer-based reporting, geographic information systems, management information and XML, as a set of building blocks for the development of convergent journalism. Therefore today's

modern journalistic organizations should function as organizations in which the prevailing knowledge management.

6 References

Bierhoff, Jan, Deuze, Mark and de Vreese, Claes (2000) 'Media innovation, professional debate and media training: a European analysis'. European Journalism Centre.

Gentry, James 'The *Orlando Sentinel*. Newspaper of the future: integrating print, television and Web'. In *Making Change*, a report for the American Society of Newspaper Editors.

Glamann, Hank 'Equipping employees for change'. *Newspaper techniques*.

Landau, George 'Newsroom as knowledge refinery'. Presentation to seminar on information technology in newsrooms, at Ifra in Darmstadt in Germany.

Northrup, Kerry 'New skill sets required for today's "multiple media" stories'.

The Naïve Ontologist and "Intelligent" Story Telling

Mícheál Mac an Airchinnigh

School of Computer Science and Statistics, University of Dublin, Trinity College Dublin, Ireland
mmacanai@cs.tcd.ie

Abstract.

Semantic Search has become, that is to say been re-discovered to be, the key theme for Intelligent Systems in 2012. No humans need apply. Semantic search, as Tim Berners-Lee predicted in his seminal little book on the World Wide Web, is the ultimate goal. After the Ultimate Goal has been obtained, and all the machines are happily communicating, what is in it for us, humans? To be sure, for us, it is to be more human! Humans tell stories, to themselves, and to other humans. Many of these stories have been, in the first instance, of an oral nature. Then they were written down. Now they appear electronically. The full spectrum of the technology of our age is used to reinvent the same old stories (7 in total). Precision in identity is sought through the formal ontology. A new story concerning Vela Peeva is revealed.

Keywords: game, keyimage, magic realism, ontology, semantic search, storytelling

1. Google's Semantic Search

Out of the blue recently came the announcement that Google was doing "*Semantic Search*" [1], reported on the Wall Street Journal [2]. We already knew that Google's digitization of the books was not only for the purpose of having electronic versions of same so that they might be "offered to the public" [3], but also that the electronic text might be searchable. Moreover, it is not the word that is sought, but the word (as used) in context. Another way of expressing this, is to say that context gives meaning to the word. Meaning is what we seek. Hence the "Seeking of meaning" is another way of talking about "Searching for semantics". But what exactly does this "Semantic Search" entail? Ultimately, for Google, it might mean that the books "can talk to each other" on many different levels and at times arrive at a "Eureka!" moment of understanding which will be "output" to the Google AI Engine? Some idea of the colossal magnitude of the project is available on Wikipedia [3].

The reference to Wikipedia, cited above [1], presents an enormous amount of information on all aspects of "Semantic Search". Of particular note, relevant to the "Island of Ireland", is the "Semantic Search Engine", SophiaSearch [4], developed in Belfast [5]. It eschews the idea of "(formal) ontology" and has just introduced the "Digital Librarian" [6]. But Wikipedia itself is a foundation

for its own“inner Semantic Search.” Specifically, there is now a new initiative: “Wikidata” which “aims to create a free knowledge base about the world that can be read and edited by humans and machines alike. It will provide data in all the languages of the Wikimedia projects, and allow for the central access to data in a similar vein as Wikimedia Commons does for multimedia files. Wikidata is proposed as a new Wikimedia hosted and maintained project.” [7, 8].

We have also become accustomed to rely on the use of a formal ontology [9] such as the CIDOC Conceptual Reference Model [10] supported by (electronic) tools such as Protégé [11].

In the context of the WWW, we know from Tim Berners-Lee, who had a dream and in the second part of that dream “Machines become capable of analyzing all the data on the Web—the content, links, and transactions between people and computers. A “Semantic Web,” which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy, and our daily lives will be handled by machines, leaving humans to provide the inspiration and intuition [12, 13].

1.1 Нека си представим?

Let us imagine that the Berners-Lee view of the “Semantic Web” has come about? Let us imagine that Google’s “Semantic Search” is available to us? What shall we (as individual) do? In this paper we propose a return (ricorso) to the human normality of individuality of the person. In other words, given all of this “newly found” technology, how shall we be (more) human? In what sense shall we be free?

Storytelling is an ancient art form, originally oral, and now electronic. The key to good storytelling lies in the use of

(i) stock phrases (“Once upon a time”),
(ii) scene setting (“They wandered into the Dark Wood”),
(iii) anticipation (“They heard footsteps behind them”,
(iv) ambiguity (“What Big Eyes you have, Granny”),

and so on. A story can be re-told to the same people. Each re-telling adds to the understanding and pleasure of the listener.

In our times, it is customary that a story be accompanied by pictures, images that invoke some key aspect of the scene, or of a character, or of an object. Pictures have always been used to “tell stories” since, at least, Stone Age Man.

][How can I find information relating to this claim?**][**

A search is required. But to do the search I normally need to use some “key words.” Ideally, a key word may be formally attached to (or be derived from) a lemma (which is defined to be the *canonical form*, *dictionary form*, or *citation form* of a set of words (headword)) [14]. One needs to adopt a specific lemmatization [15] and harmonise it with the constructed ontology that I have been using for years, the CIDOC CRM 5.0.2 [16], processed by Protégé (now at version 4.2) [11].

For the record (of the intelligent story) we may wish to exhibit (and subsequently ontologize) the plan of the house of a character, say, Todor Stoinmenov Peshterski [17], who once lived in Velingrad (see Fig.1). In the figure shown, there is a photograph of what seems to be an elderly couple from the village of Draginovo. One assumes that the house shown, built in 1890, belongs to the man on the left, by the name of Yavor(?) Karabelov. Ontologization of this picture will provide us with a straightforward negation, that the house shown is not that of Peshterski.

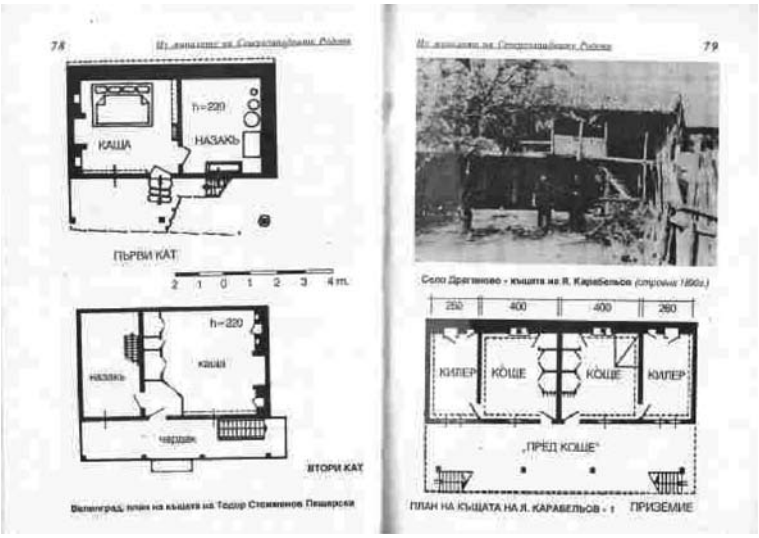


Figure 1: Велинград, план на къщата на Тодор Стоименов Пещерски

There have been, and there are, many “electronic” storytelling software platforms, such as LE-Story [18], SAGE [19] and KidPad [20]. Undoubtedly, there will be many more in future. I am not concerned with such software platforms. My interest is in the nature and structure of the story to be told. It is said that the plays of Shakespeare cover all of the stories that one might want to tell. That is to say, Shakespeare has covered all the storytelling categories. What are they?

Not remembering the original source, one seeks to learn in the usual way in 2012 [21].

There are only **7** stories :

1. Tragedy: Hero with a fatal flaw meets tragic end. Macbeth [22] or Madame Bovary.

2. Comedy: Not necessary laugh-out-loud, but always with a happy ending, typically of romantic fulfilment, as in Jane Austen.

3. Overcoming the Monster: As in Frankenstein or 'Jaws'. Its psychological appeal is obvious and eternal.

4. Voyage and Return: Booker argues that stories as diverse as Alice in Wonderland and H G Wells' The Time Machine and Coleridge's The Rime of the Ancient Mariner follow the same archetypal structure of personal development through leaving, then returning home.

5. Quest: Whether the quest is for a holy grail, a whale, or a kidnapped child it is the plot that links a lot of the most popular fiction. The quest plot links Lords of the Rings with Moby Dick and a thousand others in between.

6. Rags to Riches: The riches in question can be literal or metaphoric. See Cinderella, David Copperfield, Pygmalion.

7. Rebirth: The 'rebirth' plot - where a central character suddenly finds a new reason for living - can be seen in A Christmas Carol, It's a Wonderful Life, Crime and Punishment and Peer Gynt.

Naturally, there must be Bulgarian language categories to cover the same ground. What are they? "Google translate" provides good possibilities: трагедия, комедия, преодоляването на чудовището, пътуването и връщане, търсене, дрипи до богатство, прераждане. Now we need to give corresponding examples, but from Bulgarian culture. For example, what is the Bulgarian "equivalent" to Macbeth? Maybe Хан Аспарȕх [23]? As an aside, it is noteworthy that the Bulgarian Wikipedia often marks the stress syllable! Here is not the place to provide all these detailled inter-cultural correspondents. Google's Semantic Search engine will do that for us, if it is really "intelligent".

We need to have an intelligent storyteller, one who fits into a category, for our times. The rebirth category (**7** above) suggests… renewal… Будител! One who awakens, enlightens, inspires, … In principle, the Buditel is a(n) enlightener, a(n) (intelligent) story teller, one who looks to the future, remembering the past that can never be changed! Let us see how to play the role? We will need characters for the story. We need to enhance our continuing theme of Bulgarian Art/Culture. We will pick from the dead, and from the living. We will begin with a man (Тома ПЕТРОВ) who benefitted from the Bulgarian Enlightenment and continue with a woman Илона ЗАХАРИЕВА, a painter and former curator of Kyustendil Art Gallery (Кюстендил), the centre of the Maistora Collection.

1.2 Encoding the painter/painting "Тома ПЕТРОВ 1908-1972"

I am looking at a self-portrait of Toma Petrov (Тома ПЕТРОВ) as I write [24]. He is wearing a hat and smoking a cigarette. He wears a large overcoat and around his neck is a beautiful green scarf. The background to the self-portrait is mottled bright red, giving a purplish effect. One wonders what image springs to mind for the reader? At first the image was deliberately omitted in the belief that the reader would be able to find that self-portrait online and verify that it is indeed Toma Petrov! Having failed to find such an image online, I have included it here (in black and white):

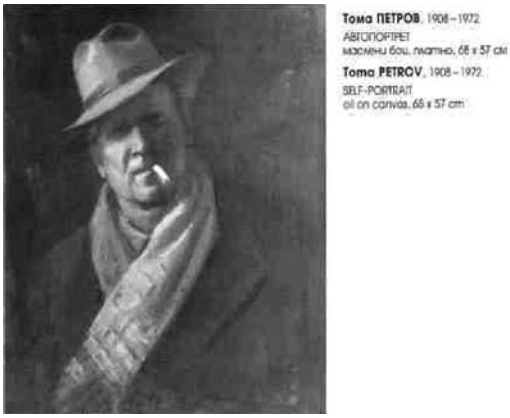


Figure 2 Тома ПЕТРОВ 1908–1972

Based on past experience of using Protégé and having checked the Bulgarian Wikipedia, it seems reasonable to encode Toma Petrov in the form:

E21 Person: ПЕТРОВ_Тома_1908-1972

E70 Thing: http://www.flickr.com/photos/mihalorel/7086955705/

E78 Collection: Sofia Art Gallery

E53 Place: СОФИЙСКА ГРАДСКА ХУДОЖЕСТВЕНА ГАЛЕРИЯ

where we are still using the Conceptual Reference Model (CRM) Ontological framework, with the aid of Stanford's Protégé 4.2. Note that there is a reasonably good representation of the self portrait recorded as E70 and located on Flickr.

There are 43 persons identified as Toma Petrov on the Bulgarian Wikipedia [25]. Of those, 8 persons are listed but have no corresponding entries. On the corresponding page of the English Wikipedia [26] there are around 60 persons identified by the patronymic Petrov, none of whom is the artist in question. One obvious interpretation/hypothesis is that our artist is unknown, or of no importance, in the realm of Bulgarian Art?

For the woman, the (former) Curator and Painter, we choose her own visual identification

Figure 3 Ilona Zaharieva

and encode (it/her) ontologically as

E21_Person: ЗАХАРИЕВА_Илона_р.1950
E22_Man-Made_Object: Автопортрет_ЗАХАРИЕВА_Илона_2000
E35_Title: АВТОБИОГРАФИЧНА КВАДРОКАЛУЦИЯ, 2000
E78 Collection: Sofia Art Gallery
E53 Place: СОФИЙСКА ГРАДСКА ХУДОЖЕСТВЕНА ГАЛЕРИЯ

This keyimage, shown above, is specifically chosen for the "Intelligent Story", precisely because the author has met the person portrayed: Илона Захариева, former director of the Art Gallery in Kyustendil (Кюстендил) and as mentioned already, Kyustendil is chosen specifically because it uniquely associated with Vladimir Dimitrov — Maistora. In addition she was born in the same year as Mihal Orela.

1.3 Groups of Painters, Schools of Painting

One obvious easily accessible starting point for the general collective of public knowledge about Bulgarian artists is in Wikipedia [27]. Another might be the great variety of books on the subject, such as an historical scholarly account of Bulgarian Art in the context of a certain period such as 1904-1912 [28] or contained in certain archives such as that of the Bulgarian Academy of Sciences (BAS) [29]. A third popular source is the catalogues and books associated with either permanent exhibitions or special exhibitions. One can ontologise one's favourite collection of painters in the form:

E74_Group: Bulgarian_Painters [same as]

E74_Group: Български_художници

Again, for the purpose of the (Intelligent) Story Telling one has recourse to the personal (re-)sources. The *early period* of "Тома ПЕТРОВ 1908-1972," is well summed up in Milena Georgieva's book (Sofia, 2008), "South Slav Dialogues in

Modernism." Therein is mention of "The Exhibition of Serbian Artists of *Lada* Union at Trupko's Gallery in Sofia (1911)" p.165.

One will be aware of the significance of Lada (Лада) "as vehicle" and even more so with Union ("Съюз"). However, although one might speculate about the possibility of "*Lada* Union" as a car-sponsoring organization, a more plausible (?) meaning is to be derived from Lada as Goddess [30]. Given the many possible interpretations, we shall decide to enter Лада (Lada) into the ontology with preferred reference to the Bulgarian Wikipedia entry [31].

E21_Person: Лада
Now we have ontologically stated that "our Lada" is a (female) person and not the car! Of course, there must be many more assertions to ground "our Lada"'s (presumed original) existence and reality, since she is a real character in an Intelligent Story. We will introduce Lada into our Intelligent Story later below.

The names of people are always of prime interest in storytelling. In Georgieva's book there is mention of the place, the Art Gallery, where "Trupko" [Тръпко] hosted the Exhibition. Fortune would have it that Google's "semantic search" [in Bulgarian] returns a unique Wikipedia page for Тръпко Василев [32]. The page is unique. There is currently no electronic equivalent in any other natural language? From the bibliographic information given, we also know that Тръпко was born in Личища [33].

E21_Person: Тръпко [same as] **E21_Person**: Тръпко Василев [same as] …
E53 Place: Личища (originally part of …)(now in Greece)(close to Ohrid)…

But here it is enough to get started with the Игра (game/play)! It is a happy fact that there is just one Bulgarian word for the "Play." We can use it ambiguously to describe the story we are telling, as a play in, lets us say, 3 Acts? Since the paper limit is about 10 pages or so… We can set down the Play structure in the frame of the "Digital re-Discovery of Culture (DrDC) Game of Inquiry!" [34].

2. Let us imagine that the (Game/Play) defines the Culture?

We begin, as we must, with the identification of the player, or storyteller, or reader, or whomever. Here is the storyteller identified by:
OpenID: https://profiles.google.com/Mihal.Orel/about [8, p51]

2.1 Prelude

Lada is so ancient that there could not be an image of her that would survive into modern times? Could she be the equivalent of the figure Eve? Might not her Spirit

re-appear again and again? Why did Ilona Zaharieva become, not only an artist in her own right, but take on the role of Director of the Art Gallery in Kyustendil?

2.2 The Play's the Thing: Game as Research Method

The idea of using Игра (play/game) to explore the issues of Culture was presented in an early paper on the "Digital re-discovery of Culture" and the Game of Inquiry/Identity [34]. There are 5 basic elements: the Backstory text, the trinity of Web pages, the visuality of Keyimages including video, the End Goal, and the Keywords [35]. At the time, from experience, it was felt that the game needed focus and direction. The keyword was intended as a strong direct hint to the player. Let us sketch a basic game template?

[1] *The Backstory (after Menard [36])*

When I was a young boy, around the age of 12, I visited Velingrad with my parents. I remember well the old house of Todor Peshterski, a very close friend of my father. I remember going into the storeroom secretly while the adults were busy, drinking some rakia and/or ayran. There was lots of old stuff, especially things that I knew to be Ottoman. But in one corner there was a picture of Peshterski, all dressed up in strange coloured clothes like the haiduk revolutionary, Panayot Hitov [37]. It was a coloured painting about my size. I went closer to take a look. I was mesmerized. I do not know how long I stared at the portrait. My mother and Maika Peshterski had come to find me. They said that I looked just like a statue, staring at the painting. That was the day I knew I would become a painter.

[2] *Web Pages* (3)

1) http://bg.wikipedia.org/wiki/Вела_Пеева [20120419]
2) http://bg.wikipedia.org/wiki/Панайот_Хитов [20120419]
3) http://bg.wikipedia.org/wiki/Пещерски [20120419]

[3] *Keyimage(s)*

Every image is semantically (over)loaded, being layered with many meanings, each of which is revealed according to the cultural background of the beholder and the circumstance of the time and place of such beholding. The image is decoded by the Eye of the Beholder depending upon the circumstances at the time. For this game, the following keyimage is chosen:

[4] *Goal*: When and where did Toma Petrov meet Vela Peeva?

[5] *Keywords*: Vela Peeva, Chepino, Ladzhene, Kamenitsa, Kleptuza

The play/game is won/finished when the player achieves internal personal psychological closure. There may be more than one ending/solution to the play/game depending on the culture of the player.

2.3 Postlude

There is no evidence to suggest that Lada participated in the painting frenzy in Lascaux [38]. It is good to remember that we have been playing with words, with images, with history. Characters have been brought together, some real, some fictional. Since there are many images involved, one needs to be able to connect them with the corresponding Ontology. Currently, the author uses KeyNote (the equivalent of PowerPoint) and Flickr for this purpose. In the Protégé ontology, the comment section is used to host the links to the images concerned. Finally, it must be understood that we have been playing a game in the strict sense. Much of the fun is obtained by finding the deliberate contradictions and impossibilities. It is Category 7: Rebirth, rediscovery, re-enlightenment. Ultimately, one needs to know how Velingrad came to be!

3 Acknowledgements

A special thanx are due to Михал Орела who provided much of the Bulgarian material used in the Intelligent Story Telling.

References

1. Wikipedia editors *Semantic search,* http://en.wikipedia.org/wiki/Semantic_search. 2012.
2. Efrati, A. *Google Gives Search a Refresh*. 2012 [cited 2012 March 29]; Available from: http://online.wsj.com/article/SB10001424052702304459804577281842851136290.html.
3. Wikipedia Editors *Google Books Library Project*. Wikipedia, 2012.
4. SophiaSearch. 2012; Available from: http://www.sophiasearch.com/.
5. Wikipedia Editors *Sophia Search Limited*. 2012.
6. SyncNI *Sophia Search Discovers US Market*. 2012.
7. Wikimedia. *Wikidata*. 2012 [cited 2012 April 13]; Available from: http://meta.wikimedia.org/wiki/Wikidata.
8. Perez, S. *Wikipedia's Next Big Thing: Wikidata, A Machine-Readable, User-Editable Database Funded By Google, Paul Allen And Others*. 2012 [cited 2012 April 13]; Available from: http://techcrunch.com/2012/03/30/ wikipedias-next-big-thing-wikidata-a-machine-readable-user-editable-database-funded-by-google-paul-allen-and-others/.
9. Wikipedia editors *Ontology (information science)*. 2012. **2011**.
10. *The CIDOC Conceptual Reference Model*. [cited 2011 Feb 27]; Available from: http://www.cidoc-crm.org/official_release_cidoc.html.

11. Stanford Center for Biomedical Informatics Research. *Protŭgŭ 4.2*. 2012 [cited 2012 April 5]; Available from: http://protege.stanford.edu/.
12. Berners-Lee, T. and M. Fischetti, *Weaving the Web : the original design and ultimate destiny of the World Wide Web by its inventor*. 1st pbk. ed2000, New York: HarperCollins Publishers. ix, 246 p.
13. W3C. *Semantic Web*. [cited 2012 April 13]; Available from: http://www.w3.org/standards/semanticweb/.
14. Wikipedia editors *Lemma (morphology)*. 2010.
15. Wikipedia Editors *Lemmatisation*. 2012.
16. *Erlangen CRM / OWL 2010-10-01 based on CIDOC CRM 5.0.2 January 2010*. 2010 [cited 2011 Jan 28]; Available from: http://erlangen-crm.org/current-version.
17. Арнаудов, А., *Из миналото на северозападните родопи,*1995, Пазарджик: ИК „Беллопринт“.
18. Yu-Lin Chu, Chun-Hung Lu, and W.-L. Hsu *LE-Story: An Intelligent Storytelling Environment for Chinese Learning*. 2009.
19. MIT. *SAGE (Storyteller Agent Generation Environment)*. 2012 [cited 2012 April 4]; Available from: http://xenia.media.mit.edu/~marinau/Sage/.
20. University of Maryland. *KidPad*. 1998 [cited 2012 April 4]; Available from: http://www.cs.umd.edu/hcil/kiddesign/kidpad.shtml.
21. Haig, M. *The Seven Stories That Rule the World*. 2008 [cited 2012 April 16]; Available from: http://us.penguingroup.com/static/html/blogs/ seven-stories-rule-world-matt-haig.
22. Wikipedia Editors *Макбет*. 2012.
23. Wikipedia editors *Asparuk of Bulgaria*. 2012.
24. Petrov, T., *SELF-PORTRAIT*, 1908-1972, “SELF-PORTRAIT, The visible image and the hidden meaning”, Sofia Art Gallery, 2008: Sofia.
25. Уикипедия. *Петров*. 2012 [cited 2012 March 26]; Available from: FinalPaper.doc.
26. Wikipedia. *Petrov (surname)*. [cited 2012 March 26]; Available from: http://en.wikipedia.org/wiki/Petrov_(surname).
27. Уикипедия. *Категория:Български_художници*. 2012 [cited 2012 Аврил 14]; Available from: FinalPaper.doc.
28. Georgieva, M., *South Slav Dialogues in Modernism* 2008, Sofia: Bulgarian Academy of Sciences.
29. BAS. *Музей “Стара София” - Художествена колекция*. [cited 2012 2012-03-18]; Available from: http://oldsofiaart.cl.bas.bg/.
30. Wikipedia Editors *Lada (goddess)*. 2012.
31. Уикипедия *Лада*. 2012.
32. Wikipedia Editors *Тръпко Василев*. 2012.
33. Wikipedia Editors *Личища*. 2012.
34. Mac an Airchinnigh, M., K. Sotirova, and Y. Tonta, *Digital Re-discovery of Culture Game of Inquiry & The Physicality of Soul, 2006.* Review of the National Center for Digitization., 2006: p. 19-37.
35. Mac an Airchinnigh, M., *Digital re-discovery of Culture and the Game of Inquiry/Identity*.
36. Wikipedia editors *Luis. Pierre Menard, Author of the Quixote. 1939*. 2012.
37. Wikimedia Commons. *File:Panayot Hitov Photo.jpg*. [cited 2012 April 18]; Available from: http://commons.wikimedia.org/wiki/File:Panayot_Hitov_Photo.jpg.
38. Wikipedia Editors *Lascaux*. 2012.

AnatOM – An intelligent software program for semi-automatic mapping and merging of anatomy ontologies

Peter Petrov[1], Nikolay Natchev[2,3], Dimitar Vassilev[4], Milko Krachounov[1], Maria Nisheva[1], Ognyan Kulev[1]

1 Sofia University "St. Kliment Ohridski", Faculty of Mathematics and Informatics, 5 James Bourchier Blvd., Sofia 1164, Bulgaria
2 Vienna University, Dep. Int. Biology, Austria, Vienna 1090, Althanstrasse 14
3 Sofia University "St. Kliment Ohridski", Faculty of Biology, Dept. Zoology, 8 Dragan Tzankov Blvd., Sofia 1164, Bulgaria
4 AgroBioinstitute, Bioinformatics Group; 8 Dragan Tzankov Blvd , Sofia 1164, Bulgaria.

Abstract. Anatomy ontologies (AOs) of various categories of species are nowadays publicly available both in scientific literature and on the web. Having various species-specific AOs is valuable when it comes to text searching or text mining in the context of a particular biological category (species, genus, family, etc.). Currently though, these ontologies do not have the necessary means to aid researchers in performing various tasks like cross-species text searching and mining, or translational research tasks.

Solving such challenges is done through mediating AOs. Ontology mediation is an area of research in the field of ontology engineering. It has been explored by the ontology mapping/merging/matching (OM) community ever since the emergence of the idea for the Semantic Web.

In this work, we view AOs as directed acyclic graphs (DAGs). To mediate (map and then merge) two or more species-specific (input) anatomy ontologies means: to draw cross-species semantic links (is_a, part_of, synonymy, others) between the DAGs representing the input ontologies; (ii) whenever possible to merge the nodes and the edges from the input ontologies, thus generating as output more general (output) ontology (called in this work a super-ontology). The only informal requirement for the super-ontology is that it makes sense from the points of view of the two input species-specific AOs.

The informal requirement to make sense and the knowledge representation models (ontologies, DAGs) employed turn that problem into a task for building an intelligent software system for integrating anatomy ontologies. Our system is called AnatOM and is able to map input AOs and to merge them into a super-ontology. The name AnatOM is an abbreviation from Anatomical Ontologies Merger.

In this paper we give an outline of the models and the algorithmic procedures that we use for integrating/mediating AOs and a detailed description of the software program AnatOM which implements them.

Keywords: anatomy ontology, mapping, merging, software, graph models

1. Introduction

Ontology mediation in general is a broad field of research within the area of ontology engineering. It is well-known that this field has been an important, popular, and well-studied subject ever since the emergence of the idea for the Semantic Web [15], as it quickly became apparent that different authors, research and business organizations will ever hardly agree on a single common set of structured vocabularies (ontologies) to be used uniformly for annotating data over the web.

Some authors, as e.g. Jos de Bruijn et al. in [14], define ontology mediation as a combined term which denotes all of the processes of ontology mapping, merging and aligning. As per them, 1) ontology mapping is mainly concerned with the representation of the semantic links that exist either implicitly or explicitly between different ontologies which share similar or identical domains of study; 2) ontology aligning is the process of automatic or semi-automatic discovery of those links; 3) ontology merging is the process of unifying or uniting the knowledge contained in two or more different ontologies under the hood of a single more general ontology.

Other authors, as e.g. Euzenat and Shvaiko in [16], use different terminology for the processes related to ontology integration. As they define it, 1) ontology matching is the process of finding relationships or correspondences between entities of several different ontologies; 2) ontology alignment is the set of correspondences between two or more (in case of multiple matching) ontologies (and, as the authors note, this term is adopted by analogy with the sequence alignment term well known from the molecular biology); 3) ontology mapping is the oriented, or directed, version of an alignment and it maps the entities of one ontology to at most one entity of another ontology (which, as the authors of this terminology claim, complies with the mathematical definition of a mapping instead of that of a general relation); 4) ontology merging is the creation of a new ontology from two, possibly overlapping, source ontologies and through this process the initial ontologies remain unaltered while the merged ontology is supposed to contain the knowledge of the initial ontologies.

In the current paper, and in the other papers of ours ([1], [2], [3]) we adopt the terminology of Jos de Bruijn [14] as we find it simpler and better suited for the purposes of our research. In addition to that, we sometimes also use the term ontology integration as a synonym of ontology mediation.

Anatomical data is nowadays publicly and freely available on the web. Usually though, this data is segmented and scattered across different AOs and databases which are species-specific/species-centric. Throughout this paper, the abbreviation AO (adopted from [17]) is used to denote a species-specific anatomy ontology. The segmentation of AO data, hinders a lot of biologically meaningful

operations which (if the data gets integrated/mediated) could be performed on that data.

Often, in biology and biomedicine, it is the case that experimental results about a particular model organism A (e.g. mus musculus), may turn out more general and thus applicable to another organism B (e.g. danio rerio or xenopus), or at least may provide valuable insights about the design of new biological experiments to be performed with that other organism B. These are some samples of translational research related problems and operations. The current state of the AO data available doesn't allow to directly perform such intelligent cross-species text searching and text mining queries. Achieving that is one of the motivation items for trying to map and merge AOs of different biological categories (species, genera, families, etc.).

Individual AOs are quite useful for pulling data from separate databases, dedicated to a particular biological category of organisms, but integrative queries which span across multiple databases or across distinct species is still not well supported. That is so because each database uses its own ontology that is designed and implemented according to different strategies, principles and purposes. There is still a large shortage of cross-ontology links between AOs, and almost a lack of links from anatomy to other domains such as e.g. genotype and phenotype. Solving these problems (which are also very well noted in [17]) is another motivation item for trying to mediate AOs from different biological categories.

There's currently also a lack of decent mechanisms for querying AO data between human on one side, and various model and non-model organisms on the other side, due to the multiple differences in the terminologies describing their anatomies [17]. This also poses a significant problem on translational research, e.g. on translating model organism lab research data to the topic of human health. Overcoming those problems is another motivation item for trying to solve the problem of mapping and merging AOs of different species.

Two recently published studies ([13], [17]), presented Uberon – a large-scale project for integrating the available species-specific AOs of various biological categories from the Animalia kingdom into a single general species-neutral anatomy ontology. As it turns out, Uberon is a multiple phase project which, as noted by its authors, started with automatic seeding of the initial version of the Uberon ontology (done mainly through lexical/direct matching from the existing species-specific AOs); went on with rigorous and careful manual curation performed by several knowledgeable experts; largely utilized computational reasoning. These three main methods were applied iteratively several times to arrive at the current state (as described in [17]) of the Uberon ontology.

2. Models and Procedures

We base our approach for AO mediation on two natural graph models and three concrete algorithmic procedures. These models and procedures have already been described from different viewpoints in [1], [2] and [3] but for the sake of clarity, we also provide a brief outline of them here.

For solving the task at hand the mathematical apparatus of graphs and graph theory is used. We assume that two AOs are given in the form of OBO files [4]. We view these given OBO-encoded ontologies as two directed acyclic graphs (DAGs) with their nodes denoting anatomical concepts and their edges denoting relations between these concepts (is_a, part_of, others).

Despite the several different terminologies, popular among the OM community, which we noted in the introduction of this paper, and which all pertain to the integration of ontologies, it is our understanding that within its core, the problem of integrating AOs is essentially a graph theoretical problem which could be tackled with the apparatus of graph theory and by utilizing DAGs in particular.

For achieving an acceptable level of intelligence while mapping the AOs, several external knowledge sources are interrogated or consulted – the FMA [6] (a detailed ontology of the human anatomy), the UMLS [5] (an exhaustive database of all kinds of biomedical information in several natural languages) and WordNet [7, 8] (a large lexical database of the English language in general). The usefulness of these three external knowledge sources, for the purposes of mapping and merging AOs of different species, has been described in [12].

On the two given AOs (on their DAGs), we run three distinct algorithmic procedures (DM, SMP, CMP) [3]. By doing this, we establish various cross-ontology links between their nodes thus mapping the two input ontologies onto each other. The output of this process is a graph model that we call ‚the two input ontologies mapped onto each other‘ [1]. Another process then runs which takes this model as input, promotes certain nodes to output nodes, processes the inner-ontology links given, and also the cross-ontology links established (while mapping the two input ontologies), and ultimately generates a second model – a single output ontology which we call ‚the super-ontology‘ [1]; that output super-ontology is also a DAG.

The DM procedure that we use is a typical lexical/direct mapping procedure which scans the two input ontologies and looks for textual identities between their terms; the lexical mapping approach is popular and widely used within the field of ontology mediation. The SMP procedure utilizes the available background knowledge sources (UMLS, FMA, WordNet) in a way similar to the approach described in [18] where the two processes, performed within what we call SMP, are called anchoring and deriving relations. As defined in [18], anchoring is the

process of matching the terms from the two input ontologies to the background knowledge sources available and deriving relations is the process of discovering cross-ontology relations between the terms from the two input ontologies by looking for relations between their anchored terms in the background knowledge sources i.e. by basically transferring those to relations between the original terms. The most original part of our work is the CMP procedure, which tries to build upon the results found by DM and SMP, through the use of a probabilistic-like function which evaluates the probability that two terms from the two input ontologies are synonymous even though they haven't been detected as synonyms by either DM or SMP. The CMP procedure works by looking for certain patterns of connectivity within the DAG representing the two input ontologies mapped onto each other, and by predicting and scoring synonymy relations between parent nodes/terms based on the relations found so far (by DM and SMP) between their child nodes/terms.

The practical realization of these models and procedures is the software program AnatOM whose name is an abbreviation from Anatomical Ontologies Merger. This program's initial version has been originally implemented in Python. Later on, the program was ported to C#/.NET for various reasons – mostly for optimization purposes and for logical testing and consistency checking. In [1], the original Python program has been described only briefly. The C#/.NET program has preserved most of the main features of the original initial of AnatOM but has also added some very important enhancements (e.g. the ability to keep track of cycles in the intermediate result graph /a concept that we define below/). Our goal in this paper, is to give a more precise and a more detailed description of the C#/.NET AnatOM program and to present its main modules.

Compared to Uberon [13, 17], our work and our software could most probably be viewed as possible alternatives or additions to the first phase of the construction of the Uberon ontology (the so-called initial seeding, which we mentioned above) but probably also as a valuable overall addition to the theoretical methods, and the practical software tools and programs utilized within the construction of Uberon.

On the other hand, as noted by the authors of Uberon, their goal was not to describe generic reproducible lexical methods for generating multi-species ontologies. With AnatOM on the other hand, our goal was just that – to come up with some generic, reproducible, both lexical and semantic mapping methods and procedures, for generating multi-species ontologies (or super-ontologies as we call them). So, in that way, this is a significant difference between AnatOM and the methods and tools, utilized in the Uberon project.

3. Software implementation – AnatOM

AnatOM is an integrated intelligent software system designed and implemented to aid researchers in the processes of semi-automatic mapping and merging of AOs from different biological categories of organisms (species, genera, etc.). AnatOM has been primarily designed for working with animal AOs and so far has only been tested with such.

AnatOM is a standalone Windows application implemented in the C# language for the Microsoft.NET framework. It utilizes several MySQL databases as backend sources of biomedical knowledge. Two input ontologies are passed in to the program under the form of OBO files [4]. The three external knowledge sources (FMA, UMLS, WordNet) are represented as relational DBs managed by a MySQL RDBMS. Apart from these three DBs, the AnatOM program also uses one additional MySQL DB called just anatom.

The AnatOM program described in this paper is comprised of several main modules: (1) OBOParser.NET – a module for reading and parsing the input OBO files; (2) DataAccess – a DB communication module providing the connectivity to the several relational DBs utilized by AnatOM; (3) Graph.NET – a general-purpose graph library; (4) GraphVisualizer.NET – a visualization module; (5) a logical module which executes the three algorithmic procedures – DM, SMP and CMP; (6) an export module responsible for generating an output ontology and for exporting it in the form of an OBO file.

AnatOM doesn't use any external or third-party libraries. All modules listed here have been implemented by us for the custom purposes of the AnatOM program (even though some of them, like the Graph.NET module for example, could practically have much broader and general usage).

3.1. OBOParser.NET

The OBOParser.NET is a small library implemented by us for reading and parsing plain text OBO files and for representing them in memory in the form of useful objects modeling the OBO syntax and semantics. The OBO parser is event-based and it signals its clients when encountering various OBO structures as they are found in the OBO file being parsed. It defines classes for representing stanzas, term identifiers, term definitions, term names, term synonyms and relations between terms. All these structures are part of the OBO language as defined in [4].

As far as we're aware this is the first available C# parser for the OBO language and flat file format. For some time, we had been looking for an off-the-shelf ready-to-use third-party C# or .NET parser for the OBO language but we had been unsuccessful in finding one and so we had to implement our own parser – that is the OBOParser.NET.

The OBOParser.NET is called when loading the input AOs from OBO files; this happens from the File menu of the AnatOM program.

3.2. DataAccess

The DataAccess module contains simple classes for querying the four MySQL relational databases utilized by AnatOM – umls, fma, wordnet, and anatom. The first three databases represent the three available external knowledge sources which AnatOM communicates with, in order to produce an intelligent mapping between the input AOs that are given. The fourth database contains: (i) a set of some important curator-derived species-neutral anatomical statements (these are currently applicable mostly to vertebrate animals); (ii) all the intermediate output of the computations performed by the AnatOM program.

DataAccess is a pretty standard module allowing AnatOM to execute all kinds of SQL queries (select/insert/update/delete) and also all kinds of stored procedure calls and user-defined function calls against the four databases noted above (and against any other MySQL DB as well).

3.3. Graph.NET

The Graph.NET is a general purpose library for manipulating with graphs. It defines the classes Node, Edge and Graph and allows for storing arbitrary kinds of objects as properties of each node and of each edge of the graph. To its clients it provides methods for adding nodes, adding and removing edges, getting all edges in the graph, getting all edges between two particular nodes, getting all the parents of a given node, getting all the children of a given node, getting the start and end nodes of a given edge, counting all the nodes and all the edges in a graph, applying various useful algorithms on graphs like e.g. topological sort, Tarjan's algorithm for finding all strongly connected components in a given graph [10], Johnson's algorithm for finding all elementary cycles in a given graph [9].

The library is easily extensible and could be enriched with other useful and practical algorithms. It is also generic enough to be easily integrated into other programs and not just into the program AnatOM which it was originally designed for.

3.4. GraphVisualizer.NET

GraphVisualizer.NET is the module of AnatOM responsible for visualizing the graphs models in an aesthetically pleasing way. While working on the implementation of AnatOM, several different approaches for visualizing graphs have been examined and weighed. These approaches are well described in [11] and include the topology-shape-metrics approach, the hierarchical approach, the

visibility approach, the augmentation approach, the force-directed approach, the divide and conquer approach.

We have decided to implement the so-called force-directed approach due to its relative simplicity and intuitiveness. This approach is based on physical forces and on minimizing the kinetic energy of a given system of material points. For the purposes of AnatOM, the graph drawings it generates are quite satisfactory.

The AnatOM program presents the results of the ontology mappings applied in both tabular and graphical forms. This module gets called when the user of the AnatOM program clicks (in the table result tab) on a particular cross-ontology link which has been predicted by AnatOM through some of its three logical algorithmic procedures. When this happens, the two nodes connected by that cross-ontology link and the edges directly associated with these two nodes are being visualized in a separate graph result tab.

3.5. Logical module

This is the module which implements the three algorithmic procedures – DM, SMP, CMP [3], and is responsible for their execution. The procedures are typically run one after the other, in the order listed here. These three procedures are run by AnatOM in separate threads so that the AnatOM program can maintain its responsiveness to user initiated graphical user interface events while the logical procedures are still running.

At each moment of time, the AnatOM program maintains what we call an intermediate result graph. The initial version of this graph is composed just of the two DAGs representing the two input ontologies. Each of the three algorithmic procedures, when run, generates new cross-ontology links which as the procedure completes, are being added to the intermediate result graph. The intermediate result graph is basically the intermediate version of the output super-ontology. In that sense, the logical module's main responsibilities is running the three algorithmic procedures and keeping track of the results generated from them in the form of an intermediate result graph. At each moment of time, the user has access to this graph and can view it or edit it, through the table result and the graph result tabs mentioned above.

The logical module utilizes an important database table which we call the knowledge table. It is part of the anatom database and it contains a special set of pre-built (pre-inserted) species-neutral anatomical statements and their truth values. For generating these statements we have consulted a knowledgeable human expert in anatomy. When applying DM, SMP and CMP, the logical module consults this knowledge table and, if the prediction which is about to be made, contradicts with any of the statements there, that prediction is not presented to the end user for manual curation or for any other user action. In that way, it can

be said that the knowledge table takes priority over predictions automatically generated through DM, SMP, and CMP.

The logical module can be started from the Action menu of the AnatOM program.

3.6. Export module

The export module is the one which takes the intermediate result graph and generates the output super-ontology in the form of an OBO file. It can be started from the Result menu of the AnatOM program. In a way, it can be viewed as a reverse to the OBOParser.NET library which takes the input OBO files and loads them into the program memory. Similarly (but in a way reverse to that), the Export module takes the current in-memory representation (the intermediate result graph), and generates from it an output OBO file (which we call the output super-ontology).

An important condition for generating an output super-ontology, is the one that the intermediate result graph needs to be acyclic, in order to be called an ontology, and in order to allow AnatOM to export it to an OBO file. The AnatOM program provides currently two means for tracking cycles in the intermediate result graph – counting them, and exporting them. When counting the cycles, the AnatOM program displays for each cross-ontology edge (in the table result tab), the count of the cycles that edge takes part in. This can turn the user's attention to those edges which are "most wrongly" predicted and which "cause most cycles" (an example for such a prediction could be that "heart is synonym of eye"). When exporting the cycles, the user can take the generated cycles text file and send it a more knowledgeable anatomical expert which could then review it, and based on the cycles listed in there, could decide to reject some predicted links thus cutting some (or, in the ideal case, all) of the cycles.

The Export module can be started from the Export menu of the AnatOM program. Its run is successful only if the current intermediate result graph is acyclic. If it contains some cycles, it is all up to an anatomical expert working with the AnatOM program rejecting or changing certain predictions until the intermediate result graph contains no more cycles, and so an output super-ontology can be successfully exported.

4. Sources of intelligence

The AnatOM program relies on two distinct sources of intelligence which enable it to semi-automatically come up with some biologically meaningful mappings and a final merging of two input AOs.

The first one is represented by the available background knowledge sources (utilized by AnatOM through the execution of the SMP procedure). As noted,

currently three knowledge sources are supported by AnatOM (UMLS, FMA, WordNet) but the AnatOM program could easily be extended with more, if adding those is anticipated or shown to be practically useful.

The second one is the knowledge table which was also already mentioned above and whose expert anatomical statements are applied after the execution of any of the DM, SMP and CMP procedures. By adding more statements to the knowledge table, we can practically train the AnatOM program and make it generate better predictions with respect to biological/anatomical adequacy.

As noted, AnatOM is not a fully automatic but a semi-automatic software program. Human knowledge and expertise is still an important ingredient and it can currently be utilized by AnatOM in two ways. The first one is what we call meta manual curation, this should be done with care and by trusted anatomy experts only. It pertains to extending the knowledge table with useful anatomical statements and, as noted, can be viewed essentially as training the program. The second one is the standard manual curation done by a user of AnatOM working on two particular input AOs. We view this simply as part of the normal process of working with the program and so the knowledge generated through this process is not memorized by AnatOM and not used as any kind of training data.

5. Results and Discussion.

On Fig. 1, we show the typical flow of control of the AnatOM application. It can be seen that the user starts work by loading the two input AOs which are then read and parsed by the program thus producing the initial version of the intermediate result graph (RG1). The available algorithmic procedures (DM, SMP, CMP) are then run by the user in the order specified here. After the termination of each procedure, a set of manual curation actions (accept certain prediction, reject certain prediction, change a prediction's direction, change a prediction's relation type) can be applied by the user in the order depicted on the diagram. The algorithmic procedures and the manual curation actions are both represented by unidirectional solid arrows. At each moment of time, the program keeps in memory the current state of the result graph which gets modified by the algorithmic procedures and by the manual curation actions. The various versions of the result graph are represented by the rectangles labeled RGN, for N=1, 2,..., 7. It can be seen that at almost each moment of time, the user can decide to save the current state of the result graph (and, even though that is not depicted on the picture, also the set of the decisions/curations he/she has made so far) to the DB. Later on, the user may decide to move on from where he/she left off by loading back the saved state of the result graph (and the set of decisions/curations he/she has made before) from the DB. These save/load operations are depicted by bidirectional dashed arrows. Once the user is done working with the application,

he/she has the option to export the final version of the result graph (RG7) to an OBO file representing the output super-ontology.

On Fig.2, we show the different modules of the AnatOM application and how they are related to or dependent on each other. The input/output OBO files are also shown on this diagram linked with dashed lines to the modules which read/write them. Each of the solid arrows depicted points from module A to module B if and only if module A is using/calling module B and module B is the module used/called by A i.e. the start of the solid arrow is always at the client module and the end is at the module being used/called. For example, there's an arrow from GraphVisualizer.NET to Graph.NET which means that the module GraphVisualizer.NET uses/calls the module Graph.NET.

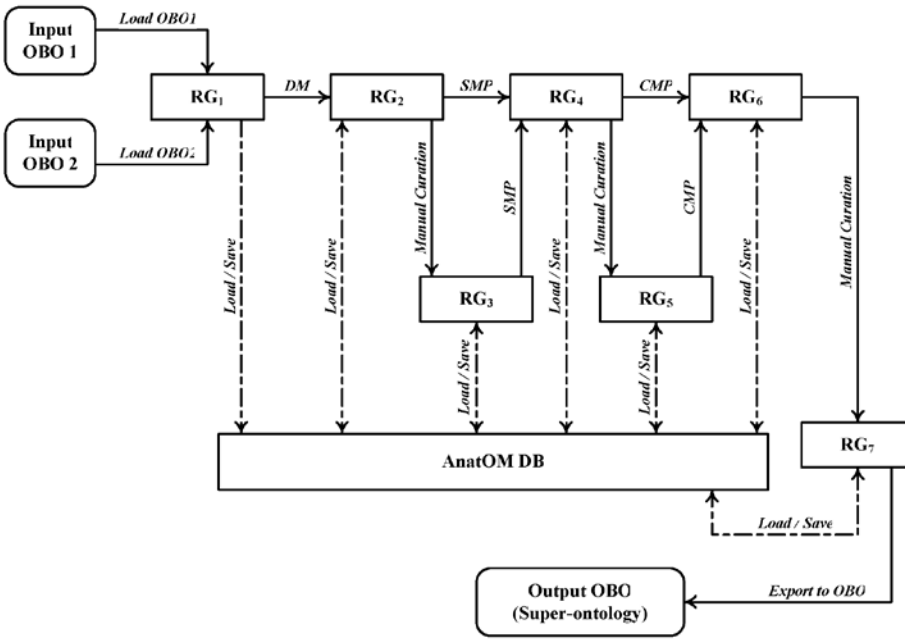


Fig. 1 Typical program flow of the AnatOM application

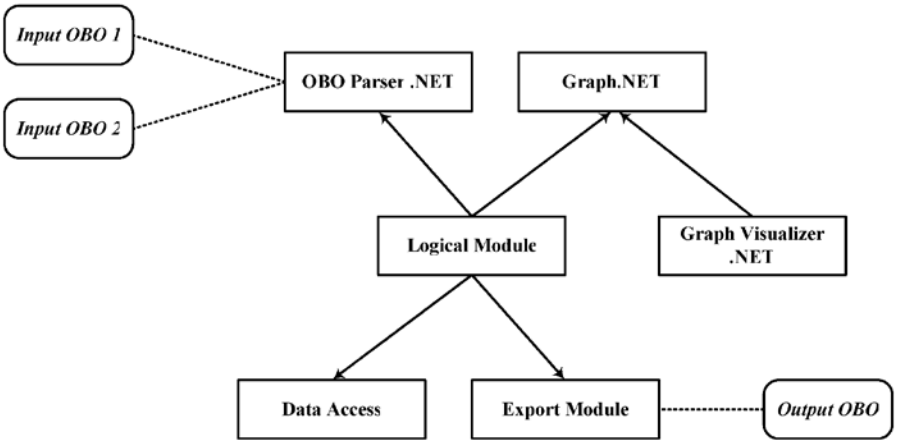


Fig. 2 Module interdependencies within the AnatOM application

AnatOM has been originally designed to work with (at least) three distinct external knowledge sources – UMLS, FMA, WordNet. Through experimentation though, we have noticed that, as WordNet is not specific to biology, it introduces some serious semantic issues when used. So in the current version of AnatOM we have largely excluded it from the tests and experiments that we have performed. On Table 1, we show the CMP synonymy predictions generated by AnatOM (when using UMLS and FMA only). These were then manually assessed by a knowledgeable anatomy expert. The experiments were performed by doing 3 pairwise mappings using 3 publicly available AOs – those of zebrafish (Danio rerio – a species), of house mouse (Mus musculus – also a species) and of a popular kind of frog (Xenopus – a genus).

Table 1: Predictions and their assessments from the 3 pairwise mappings performed.

CMP Mapping	Total CMP predictions	Precise Match. Indeed a synonymy relation. (1)	Imprecise match. Not a synonymy but some other relation. (2)	Wrong. No match at all. (3)
Mus – Danio	660 (100%)	131 (19.85%)	416 (63.03%)	113 (17.12%)
Mus – Xenopus	562 (100%)	137 (24.38%)	389 (69.22%)	36 (6.40%)
Danio – Xenopus	583 (100%)	152 (26.07%)	286 (49.06%)	145 (24.87%)

From analyzing this and similar results, we noticed that the actual count of the true predictions (1) seems satisfactory (around 20%–25% of all predictions),

provided the CMP procedure is totally automatic and relies only on internal patterns of connectivity in the intermediate result graph and not on any external knowledge directly. Also, it can be seen that the highest count is of those CMP predictions which were assessed as (2). That is so mostly due to the fact that the different AOs have different levels of detail and granularity and that the CMP procedure applied cannot completely handle these discrepancies. So most terms were predicted as synonyms even though, after closer inspection, they are actually related by another relation (like is_a or part_of). Finally, the count of the wrong predictions seems quite volatile and largely dependent on the exact comparison that is performed (Mus-Danio, Mus-Xenopus or Danio-Xenopus). Still, the percentages in (3) seem low enough to be satisfactory. It should be noted here that all the presented values are based on a fully automatic generation of the CMP predictions. Further work with the AnatOM program from an anatomist or a biologist in general, would make the whole process semi-automatic but would surely also largely improve the results achieved and presented in the table here.

From a biological point of view, we have to discuss some topics in an attempt to explain the results from our merging of AOs by using AnatOM.

First, we have to stress that the input AOs are not homogenous with respect to their contents. Some include developmental stages and some are only focused on the anatomy of the adult organism. Here, we can expect certain conflicts simply because we cannot match directly a lot of embryonic terms to terms pertaining to an adult organism. An example is the relation of the term “coelom” to the term “pericardium”. In the vertebrates, analyzed in this study, we can match the two terms with the relation “pericardium is_a coelom”, but also with the relation “pericardium part_of coelom”.

Second, we should note another contradiction which originates from the conflict between different descriptions of potentially identical structures in Latin and English. Here, we had some struggle with the establishment of adequate relations between some loosely formulated terms like “cardiac muscle tissue” and some strictly defined terms like “myocardium”. One can use both terms as synonyms but strictly judged they may be related in other ways (e.g. thorough a part_of relation). This mainly depends on the meaning of the definition of the authors of the ontology and the interpretation of the analyzer who is using it. We can define the problem as the problem of the semantic load of terms. Especially hard to handle are terms that are semantically over- or under- loaded as e.g. the terms “portion of organism substance”, “portion of tissue”, “Xenopus anatomical entity”, “acellular anatomical structure”, “anatomical set”, “multi-tissue structure”, “anatomical space”, “surface structure”.

Third, there are also some problems with the definitions of certain terms within the input ontologies themselves. A good example is the statement “bulbus arteriosus part_of heart”. Actually, the formulation of the term “bulbus arteriosus”

leads to the conclusion that this structure cannot be part of the heart but has to be part of the arterial system (and possess smooth rather then cardiac musculature). Whenever such problematic definitions did not lead to formations of cycles in the intermediate result graph, we had accepted them in an attempt to keep the input AOs unaltered as much as possible.

6. Conclusions

The main conclusions which we can draw from our work are as follows. First, we managed to semi-automate the processes of mapping and merging AOs through the use of an assembly of various tools and techniques – a set of external specialized semantics-rich knowledge sources, a pre-built database of species-neutral anatomical knowledge, a manual curation phase, and adequate algorithmic software procedures. Second, the novel CMP procedure, that we developed and used in this work allowed us to detect certain true semantic cross-ontology links which were not found either by lexical matching (DM) or by consulting any of the external knowledge sources (SMP). Third, the AnatOM program developed as part of our work demonstrates that it is possible to have an integrated software system which can perform reproducible (as opposed to one-time only), intelligent mapping and merging of AOs, and which can generate a general species-neutral anatomy ontology (a super-ontology) from the AOs that are merged.

With respect to directions for further work, we can focus on certain improvements of some statistical characteristics of the CMP procedure like sensitivity and specificity. Other procedures that build upon DM and SMP (like the CMP does) can be developed and incorporated in AnatOM with very small effort. Also, more tests and experiments can be performed with mapping and merging of more than two AOs. Finally, the WordNet knowledge source, largely unused by AnatOM in its current version, can be incorporated back into the program, if the semantic issues which it introduces are solved.

Acknowledgments. This work has been partly funded by the Sofia University SRF within the "Methods and information technologies for ontology building, merging and using" Project, Contract No. 177/2012.

References

[1] Peter Petrov, Milko Krachunov, Elena Todorovska, Dimitar Vassilev (2012) An intelligent system approach for integrating anatomical ontologies, Biotechnology and Biotechnological Equipment 26(4):3173-3181.

[2] Peter Petrov, Milko Krachunov, Ognyan Kulev, Maria Nisheva, Dimitar Vassilev, (2012) Predicting and Scoring Links in Anatomical Ontology Mapping (submitted)

[3] Peter Petrov, Milko Krachunov, Ernest A.A van Ophuizen, Dimitar Vassilev, (2012) An algorithmic approach to inferring cross-ontology links while mapping anatomical ontologies, To appear in Serdica Journal of Computing, ISSN 1312-6555, vol. 6 (2012).
[4] John Day-Richter, "The OBO Flat File Format Specification, version 1.2", available online under http://www.geneontology.org/GO.format.obo-1_2.shtml
[5] Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32:267-270.
[6] Rosse C., Mejino JL., Jr. (2003) A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform.; 36(6):478–500.
[7] Miller G.A. (1995) WordNet: A Lexical Database for English. Communications of the ACM, 38(11): 39-41.
[8] Fellbaum Ch. (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 422p.
[9] Donald B. Johnson, "Finding All the Elementary Circuits of a Directed Graph", SIAM Journal on Computing. Volumne 4, Nr. 1 (1975), pp. 77-84
[10] Robert Tarjan, "Depth-first search and linear graph algorithms", SIAM Journal on Computing. Volume 1, Nr. 2 (1972), pp. 146-160
[11] Ioannis G. Tollis, Giuseppe Di Battista, Peter Eades, Roberto Tamassia, "Graph Drawing: Algorithms for the Visualization of Graphs", Prentice Hall, 1999
[12] van Ophuizen EAA, Leunissen JAM (2010) An evaluation of the performance of three semantic background knowledge sources in comparative anatomy. J. Integrative Bioinformatics 7, 124, doi:10.2390/biecoll-jib-2010-124.
[13] Haendel M, Gkoutos G. V, Lewis S, Mungall C, Editors. (2009), „Uberon: towards a comprehensive multi-species anatomy ontology", Buffalo, NY: Nature Proceedings.
[14] Jos de Bruijn et al., „Ontology Mediation, Merging, and Aligning", In: J. Davies, R. Studer, P. Warren (Eds.), Semantic Web Technologies. John Wiley and Sons, 2006, pp. 95-113.
[15] Berners-Lee, Tim (2001), „The Semantic Web", Scientific American, May 2001
[16] J. Euzenat, P. Shvaiko, Ontology Matching, Springer, Heidelberg, 2007.
[17] Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA, 2012, „Uberon, an integrative multi-species anatomy ontology", Genome Biol. 13:R5.
[18] Zharko Aleksovski, Warner ten Kate, and Frank van Harmelen. Exploiting the structure of background knowledge used in ontology matching. In Proc. 1st International Workshop on Ontology Matching (OM-2006), collocated with ISWC-2006, Athens, Georgia (USA), 2006.

Usage of E-learning for Improving of Knowledge Management in Organizations

Mufa Vesna [1], Manevska Violeta [1], Blazeska-Tabakoska Natasha [1]

[1] Faculty of Administration and Management of Information systems, University „St.Kliment Ohridski“, Partizanska bb, 7000 Bitola, R.of Macedonia

Abstract. The main goal of knowledge management is improving of the organizational knowledge to achieve an improved organizational behavior, improved organizational performance and to make better decisions. E–learning, which refers to learning which is enabled by electronic technologies, is an irreplaceable part of effective knowledge management. Every organization should use a proper motivation for learners that acquire knowledge by e-learning. Knowledge management and e-learning enable the entire process of knowledge, beginning from its creation to its reuse. This paper will be focused on the effectiveness of e-learning as a tool for creation and distribution of knowledge and will present the significant overlaps and objectives between knowledge management system and e-learning system.

Keywords: knowledge, knowledge management, e-learning, organization

1 Introduction

Knowledge, which differs from data and information, is an understanding of information, gathered from various sources and is the ability to turn information and data into effective action. Knowledge management is capable to create, explicate and share knowledge between members of an organization, in order to achieve an organizational objectives [1].

The key role of information technologies in the process of facilitating knowledge management, was a reason for their introduction into organizational practice. One of the frequently used forms of information technologies is implementing of e-learning for better managing of organizational knowledge. E-learning is customized learning, enabled by electronic technologies that supports virtual learning environment.

Knowledge management, together with e-learning, enable the entire life cycle of knowledge: creation, acquisition, refinement, storage, transfer, sharing and its use [2]. Although e-learning refers to individual learning, while knowledge management refers to organizational learning, a common problem is facilitating the learning process in organizations.

2 Knowledge Management and Organizations

Like people, who are unable to exploit the full capabilities that they possess, organizations are unable to utilize the knowledge that they possess. Knowledge management is the organizing, motivating and controlling of people, processes and systems in the organization, to ensure that organizational knowledge is effectively used [3].

Knowledge of the organization, which is situated in printed documents and electronic media, include knowledge about the best performing of duties, knowledge that is maintained by individuals and teams, who work on solving organizational problems and knowledge, incorporated in organizational processes. Organizations try to share knowledge and enable its availability to all, who require its use.

Employees should participate in knowledge management. Knowledge management, as an activity, leads to improved organizational behavior and performance and to make better decisions.

3 E-Learning in Organizations

The speed with which organizations introduce changes in their business, leads to the need for faster learning and absorbing new knowledge by employees. Knowledge management should enable effective and efficient education, so that for this purpose the most appropriate is using of e-learning.

E-learning, as "just in time" learning, should enable continuous delivery of knowledge from which employees have necessary at a given moment, in a format understandable for them. As a kind of distance learning, with the growth of the Internet, is becoming more acceptable.

The users of e-learning are consumers of prepared knowledge from an established repository and they should apply it in practice. The access to knowledge, i.e. to the content intended for e-learning is realized through web or intranet. The exchange of experiences, attitudes and opinions among employees can be done via email, chat or discussion forums. Knowledge dissemination contributes for increasing of team and individual performance [4].

E-learning is a complement to traditional ways of learning from books or CD provides meeting of new challenges imposed by the business world [5]. With this type of learning, employees are responsible for their own learning. It offers a possibility of learning from any location, at any time, with any style, with their own tempo, regardless of the progress of others. When employees are faced with, for them, easier or known section they go to the next, but when they are faced with a difficult and a new section there is no pressure in terms of time, which is necessary for its understanding.

One of the most important benefit is the economic benefit and time saving, due to the possibility for participation of employees of various kinds of seminars or training, which though they aren't maintained at their place of living or working, they have an opportunity to participate in them. In this way, organizations increase the likelihood to possess trained employees.

4 E-Learning as a Tool for Improving Knowledge Management

To be an effective tool for knowledge management, before investing in e-learning, several factors should be considered, which are associated with the value chain of e-learning shown on Figure 1. It includes organizational readiness, determination of appropriate content, determination of appropriate presentation and implementation of e-learning [6].

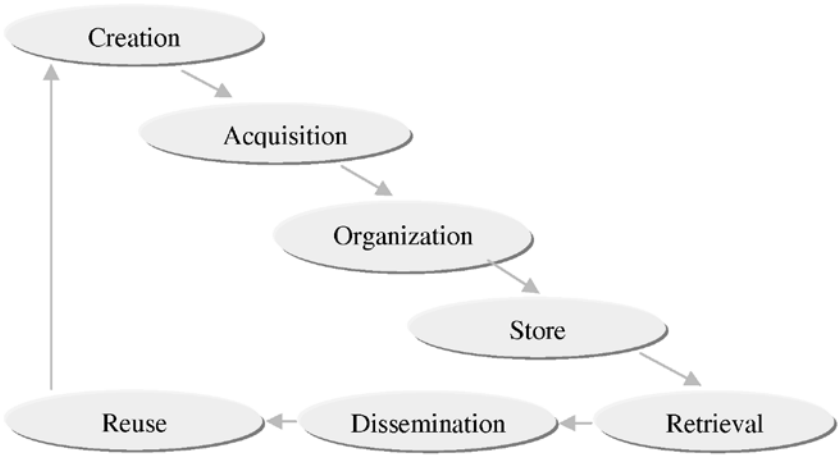


Fig. 1. Value chain of e-learning

Organizational readiness refers to, whether the organization has identified the need for knowledge, which attitude have employees in terms of knowledge management, what culture prevails in the organization, whether exist an infrastructure for knowledge management and whether the technological infrastructure is appropriate for e-learning. Firstly, should be identified requirements for knowledge and needs of knowledge management. At this stage, should be noted the views of employees regarding their readiness for self-learning, sharing or concealment of information. Employees should always be guided by the thought of sharing of knowledge, in order to overcome the barrier between employees, who possess the knowledge and employees, who have need of that kind of knowledge. The organization should establish the culture in which knowledge will be treated as property and should be aware of the advantages, offered by

e-learning [7]. The adaptation of e-learning is easier and less expensive if the infrastructure for knowledge management already exists within the organization. The infrastructure enables availability of knowledge, whenever they have a need for them, and have a role of forwarding the knowledge across organizational business processes. In terms of technology, each organization needs to check whether the existing technology is sufficient for implementation of e-learning and if it's not sufficient, then it should be extended [8].

Determination of an appropriate content depends on organizational type, needs and goals, and is characterized as explicit or implicit knowledge. The explicit knowledge is true, clearly understood, which can be more accurately documented and shared. The implicit is intuitive, internally and subconsciously understood knowledge, developed from experience and usually shared through interactive conversation, storytelling and sharing of experiences, enabled through chat, video conferencing and discussion forums. The content of e-learning should include explicit information, and e-learning system should enable sharing of implicit information. Information should be integrated, allowing their utilization in real life. The features of the content are: intelligibility, practicality, accuracy and timeliness. An enrichment of content is done through case studies, researches and if employees have access in the organizational problems.

Determination of appropriate presentation depends on the content and from a manner, which is acceptable for its understanding. E-learning must offer multiple presentation styles, from which the users can choose the most acceptable for them. Some users prefer written content, while for others, the best way to absorb a material, is through audio, video, simulations and animations. The organization of content for e-learning in audio-video presentations, contribute to retention of the material longer in the memory, due to a combination of elements that cause curiosity and interest of each individual, such as audio, video and expression of personal attitudes.

Implementation of e-learning is applied, if the previous links of the value chain are met, i.e. if the organization is prepared to accept e-learning, as a way of effectively knowledge management and if the content, intended for e-learning, and the manner of its presentation are determined. The implementation requires a planned and ready network infrastructure, to support e-learning [9]. The planned and ready network infrastructure, includes all factors that ensure smooth flow of e-learning, such as security permissions, setting up files with content and the ability to maintain video conferencing. Once the network infrastructure is ready, there are needed tools for creating applications and designing content, as well as software for running interactive lessons and for integrating all e-learning components.

A feedback from users, received as assessment or evaluation, in a form of comments or suggestions, which are sent via email or web-based electronic form,

should be used for enrichment of knowledge management system, or for re-designing of e-learning system, in order to improve their performance [10].

1.1 Motivation of the Employees for Using E-learning

If there is no appropriate motivation for knowledge acquisition by e-learning, its implementation will not have a function for improving knowledge management. In the past only motivation was a design, so suitable design was guarantee for attraction individuals and therefore, it has been made with great attention. A simple way of presented content, combined with animations and good quality of training, are resulting with motivation.

A large number of developed theories and models suggest how to increase motivation among employees. One of the most important models, is the Wlodkowski's model, which focuses on the role of motivation in the beginning of learning, when the emphasis should be placed on learning needs, during learning, when needs to be applied stimulus for learning, and on the end of learning, when the focus should be placed on competence [11]. Moshinskies's model, as newer compared to the previous, creates techniques that are complement to the intrinsic motivation of individuals who learn. According to this model, there are individuals with intrinsic motivation, which have need of little extra motivation and individuals with little or without intrinsic motivation, which have need of more extra motivation [12].

5 The Knowledge Management Life Cycle, Supported by E-learning

The knowledge management life cycle, supported by e-learning, comprising: creation, acquisition, organization, store, retrieval, dissemination and reuse of knowledge, which are shown on Figure 2.

Fig. 2. Diagram of knowledge management life cycle

The ability to create new knowledge is the heart of the organization's competitive advantage. Knowledge creation is a development of new knowledge from data, information, prior knowledge and skills or experiences [13]. The prior

knowledge, as explicit knowledge, and can be placed in the e-learning system, while the knowledge about skills and experiences, as implicit knowledge, can be transferred through online chats and video conferences. This indicates of usage of e-learning in the process of knowledge creation.

The process of acquisition of knowledge means capturing of knowledge, from various sources, such as books, journals, reports, guides, online databases and various electronic publications, thesis and dissertations. Knowledge capturing is eased by tools, developed for that purpose. Created and acquired knowledge should, pass through the process of organization that include classifying, indexing and categorization of the knowledge. Organized knowledge is set on the e-learning system for its use, by members of the organization and is stored in warehouses for its further use.

The retrieval depends on the appropriate organization of the knowledge. Tools, which are adopted in e-learning, in order to facilitate the access to knowledge, are: search engines, data mining and metadata. Data mining is used for detection of unknown patterns (forms) of existing knowledge. Knowledge dissemination, which means propagation of knowledge among members of the organization is provided by: discussion forums, talk rooms, email and video conferencing [14]. Memebers of the organization should use the knowledge that has been shared to them. The creation of new knowledge from existing, is reuse of existing knowledge.

6 Overlaps Between Knowledge Management and E-learning

The main relation between e-learning and knowledge management is that, e-learning is a tool for knowledge management, while knowledge management is a platform for e-learning [15]. This indicates on mutual support and mutual overlaps. Overlaps are related with technologies, processes, content, people and goals.

E-learning is impossible without using of Internet, information and communication technologies, while knowledge management is based on information technologies. The combination of knowledge management and e-learning requires use of databases, multimedia, Internet, intranet and extranet. Also, there are needed tools and software for implementation of e-learning. E-learning can be defined as a process, in which employees in one organization through the use of information technologies acquire knowledge, while knowledge management is the process, in which employees use information technology to explain data and information [16]. Strategies that are used to perform knowledge management, can accompany to the process of e-learning, with emphasis put on the full life cycle of the knowledge, respectively on the transmission, distribution and delivery of explicit and implicit content to individuals, groups or organizations.

The people, as key aspects of the two processes, should be enthusiastic, ambitious and always ready to learn. They constitute two groups: individuals with appropriate information, communication, interpersonal and managerial skills, responsible for designing, managing and maintaining of the knowledge management system and e-learning system and individuals involved in the process of learning, i.e. employees, which acquire new knowledge.

Knowledge management and e-learning have a main common goal - to enable access to knowledge. Overlaps, such as the development of technical infrastructure and management of implicit and explicit knowledge, are and common goals. The establishment of a collaborative environment and culture for learning is crucial for successful knowledge management.

7 Proposal for Using Knowledge Management and E-learning in Higher Education

Caused by the growth of competition in the global academic environment, universities are trying to find efficient methods for shared access to their key resources: knowledge, experience and ideas. The answer of this challenge is to use knowledge management and e-learning.

Quality education and standardization of the program structures are achieved by exploiting of the knowledge, competencies and skills of teachers. To achieve this goal, should be organized discussions and events, for exchanging of experiences and knowledge. Coordination of all teachers is a complex challenge, according to their work involvement, but there is needed a way to reach them, regardless of time and location where they are.

The solution of the above problem is the use of e-learning for establishment of collaborative platform for meeting of teachers, who are unable to achieve meeting in live. It will be used for distribution of important internal information from university to teachers. Teachers will have an opportunity to exchange ideas and practices, for successful traditional learning. The presentation of results from projects, done for e-learning as best practice, should inspire them to share their experience with colleagues. Research projects set in the e-learning system will be a subject for future analysis and further discussion among teachers.

User-friendly platform will enable easy way of use, which directly contributes to increased motivation. The opened debatable forum will serve for proposing questions, doubts, criticisms and ideas, for further development of the platform. Universities can develop its own system for e-learning or they can use ready system for e-learning. Oracle is a leading company that offers a ready system for e-learning, i.e. virtual learning environment that provides effective, manageable, integrated and extensible learning and knowledge sharing to anyone, anytime and anywhere.

8 Conclusion

If organizations decide to invest in e-learning, in order for better knowledge management, then they need to give answers to the value chain of e-learning. Overlaps between knowledge management and e-learning provide easier acceptance of e-learning by the members of the organization. However, organizations should use appropriate motivation for knowledge acquisition through e-learning. Knowledge management and e-learning, are attempting to create and share useful knowledge. Usage of e-learning in organizations, cause changes in business, through dissemination of knowledge and new information. It offers many proven benefits, of which the most important to highlight are benefits in time and money and flexibility in learning.

9 References

1. Mihalca, R., Uta, A., Andreescu, A., Intorsureanu, I.: Knowledge Management in E-Learning Systems. Revista Informatica Economica, vol. 11, pp. 60--65 (2008)
2. Islam, S., Kunifuji, S., Miura, M., Hayama, T.: Adopting Knowledge Management in an E-Learning System: Insights and Views of KM and EL Research Scholars. Knowledge Management & E-Learning: An International Journal, 375--398 (2011)
3. Khademi, M., Kabir, H., Haghshenas, M.: E-learning as a Powerful Tool for Knowledge Management. 5th International Conference on Distance Learning and Education, pp. 35--39. IACSIT Press, Singapore (2011)
4. Ponce, D.: What can E-learning Learn from Knowledge Management?. 3rd European Knowledge Management Summer School, pp. 7—12. San Sebastian, Spain (2003)
5. Cisco System: E-Learning at Cisco, www.cisco.com
6. Wild, R.H., Griggs, K.A., Downing, T.: A framework for e-learning as a tool for knowledge management. Industrial Management & Data Systems, vol. 102, pp. 371--380 (2002)
7. Davenport, T.H., DeLong, D.W., Beers, M.C.: Successful knowledge management project. Sloan Management Review, 43--57 (1998)
8. Senge, P.: The Fifth Discipline: The Art and Practice of the Learning Organization. Doubleday, New York (1990)
9. Berry, J.:Traditional training fades in favor of e-learning. InternetWeek (2000)
10. Nevo, D., Furneaux, B., Wand, Y.: Towards an Evaluation Framework for Knowledge Management Systems. Information Technology and Management, vol. 9, pp. 233--249 (2008)
11. Hodges, C.: Designing to motivate: Motivational techniques to incorporate in e-learning experiences. The Journal of Interactive Online Learning, 3--4 (2004)
12. Moshinskie, J.: How to keep e-learners from e-scraping. Performance improvement (2001)
13. Sabherwal, R., Sabherwal, S.: Knowledge Management Using Information Technology: Determinants of Impact on Firm Value. Decision Sciences, vol. 36, pp. 531--567 (2005)
14. Zhang, W., Kim, M.: Harnessing Explicit Knowledge. Journal of Economics and Behavioral Studies, 97--107 (2011)
15. Liu, Y., Wang, H.: A Comparative Study on E-learning Technologies and Products: from the East to the West. Systems Research and Behavioral Science, vol. 26, pp. 191--209 (2009)
16. Chunhua, Z.: E-learning: The New Approach for Knowledge Management (KM). International Conference on Computer Science and Software Engineering, pp. 291 – 294 (2008)

ASP.NET-based Tool for Searching in Folklore Lyrics Stored in XML Format

Dicho Shukerov[1]

[1] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski"
shukerov@fmi.uni-sofia.bg

Abstract. In this article we present the implementation of the semantic search module of an ASP.NET-based search tool in a digital library with Bulgarian folklore lyrics, stored in XML format files. We discuss the ontology file, which is the basis of this module and describes different events and folklore elements that could be found in folk songs. We will introduce search possibilities, given to us by the user interface (UI) and we will take a look ("behind the scene") how the search engine works (what steps are performed by the program to realize the search and to return the results to the user).

Keywords: ASP.NET, Search Engine, Semantic Search, OWL Ontology, Protйgй 4.1, Bulgarian Folk Songs, Folklore Lyrics, XML Files

1 Introduction

The discussed search engine is a part of the "Information Technologies for Presentation of Bulgarian Folk Songs with Music, Notes and Text in a Digital Library" project [1]. It is a web-based tool, built with ASP.NET 3.5 technology [2] and C# programming language.

The engine uses an OWL ontology file, which describes the following categories and subcategories (let us use "class" instead of "category") [3]: main classes - historical events, nature and astronomical phenomena, beliefs and rituals, labor and household items, family events and kinship ties; subclasses - war, slavery, rain, snow, etc.; their equivalent ones and members. The program takes one of these classes, gives all its subclasses and their members to the user; the user selects what to search for and presses the search button; the program takes all equivalent classes of the selected ones; does the search in lyrics for given words and returns songs that satisfy the search criteria. And because the folk songs are Bulgarian, the user interface, XML files' content and ontology's content are completely in Bulgarian. Let's see in depth ontology's structure and how the engine works.

2 The ontology

In this section we will see the structure of our ontology file – "*Folk_ontology_*

vs.owl", all classes, their equivalent ones and their members for whom we can search at the lyrics for.

Actually, the OWL file has XML structure, so we can easily parse it. In our ontology:

- Classes are stored as

```
<Declaration>
   <Class IRI="#class_name"/>
</Declaration>
```

- Members are stored as

```
<Declaration>
   <NamedIndividual IRI="#member_name"/>
</Declatration>
```

- Each relation "class-subclass" is presented as

```
<SubClassOf>
   <Class IRI="#child_class_name"/>
   <Class IRI="#parent_class_name"/>
</SubClassOf>
```

- Members of each class are presented as

```
<ClassAsserton>
   <Class IRI="#class_name"/>
   <NamedIndividual IRI="#member_name"/>
</ClassAssertion>
```

In our ontology all classes may have three forms of equivalent classes: *participant*, *synonym* and *form*.

- So, each form of class and its equivalent classes are stored as

```
<EquivalentClasses>
   <Class IRI="#class_name"/>
   <ObjectSomeValuesForm>
      <ObjectProperty IRI="#{synonym|form|participa
nt}"/>
      <ObjectOneOf>
         <NamedIndividual IRI="#member_name"/>
         [<NamedIndividual IRI="#member_name"/>…]
      </ObjectOneOf>
   </ObjectSomeValuesForm>
</EquivalentClasses>
```

Table 1 describes a part of the ontology classes and their equivalent ones that we offer to user to search in lyrics for. For each class you can see its: level; name; form, synonym, participant equivalent classes. The class hierarchy is presented in depth, i.e. parent class – all his subclasses, parent class – all his subclasses, etc.

Level	Name	Equivalent class		
		form	synonym	participant
0	историческо събитие			
1	значимо събитие			
2	война	войн, войни	сражение	войник, войници
2	въстание	въстана, въстанаха	размирици	бунтовник, четник, etc
2	освобождение	освободи, освободиха	избавление, свобода, etc	Русия, руснак, руснаци
2	поробване	робство		роб, робиня,
2	превземане	превзема, превземат, превземаха		
2	създаване	създаде, създадоха		
1	историческа местност			
1	престъпление срещу народа			
2	бесене	бесеха, бесят		
2	клане	колеха, колят		
2	побой	бият		
0	природно състояние или явление			
1	астрономическо явление			
2	залез	залезе, скри се		
2	затъмнение			
3	лунно	лунно затъмнение		
3	слънчево	слънчево затъмнение		
2	изгрев	зора, зори, изгря		
2	комета	комети		
2	метеор	метеори	падаща звезда	
2	слънцестоене			
3	зимно	зимно слънцестоене		
3	лятно	лятно слънцестоене		
1	атмосферно явление			
2	буря		виелица	
2	валеж			
3	градушка	градушки		
3	дъжд	червен дъжд, дъждец, etc		
3	сняг	снежище, снежец, etc		
3	суграшица	суграшици		
2	гръмотевица		гром, гръм, мълния, etc	
2	дъга		зуница	
2	мъгла	могла, мъглище		
2	светкавица	светкавици		
2	градушка	градушки		

2	земетресение	земетръс, земетръст		
2	лавина	лавини		
2	наводнение	наводнения		
1	природно състояние			
2	жега	жеги		
2	студ	изстина	мраз, прохлада, студенина, хлад, etc	
2	суша	засуши се	бездъждие	
0	поверия или ритуали			
1	ритуали			
2	Великден	великденски		
2	коледуване	коледа, коладе, коледо		коледари, нине
2	лазаруване			лазарки
2	сирни заговезни		пустове	кукери, кукер
2	сурвакане	сурва		
2	баене			баячка
0	битови и трудови елементи			
1	битово събитие			
2	болест	болен, разболя се	епидемия	
2	бременност	забременя	захожда, захода	
2	раждане	роди		
2	смърт	умря		
1	селски труд			
2	жътва	жънат, жъне, жънеха		жътварки, жътварка, жътвар, жътвари
2	коситба			Косач
2	овчарлък			овчар, пастир
2	оран	оре		Орач
2	чиракуване			Чирак
0	семейно събитие или роднинска връзка			
1	роднинска връзка			
2	кръвни			
3	баба	бабо, бабичко		
3	баща	бащице	татко	
3	брат	брате, братче		
3	братовчед	братовчеде, братовчедче		
3	внук	внуче, внучка		
3	дете	деца		
4	дъщеря	дъщеричка	дощерьо, щерка, щерко	
4	син	сине, синко		
3	дядо	деденце, дядка		
3	леля	лельо	тьотка	
3	майка	майко, майчице, мама, мамо		
3	сестра	сестро, сестричке		
3	чичо	кънко, чичка		
2	некръвни			
3	зет	зетко		
3	мащеха	мащехо		

3	свекър	свекъри		
3	снаха	снахо		
3	съпруг	съпруже	мъж	
3	тъща	тъщата	сватя	
1	семейна драма			
2	изневяра	изневери	измени, измяна	
1	семейно събитие			
2	бременност	забременя	захожда,	
2	годеж	сгоди		
2	кръщене	кръстилка		
2	погребение	погребаха		
2	раждане	роди		
2	сватба	зажени		
2	смърт	умря		

Table 1. Part of the classes and members used by the semantic search engine

Table 2 describes some members of classes.

Level	Name	Member
2	война	Кримска война
2	въстание	Априлско въстание
1	Историческа местност	Плевен, Шипка, Батак, Копривщица

Table 2. Some members of classes

The ontology can be edited anytime you want. The recommended software you should use is Protйgй [4] version 4.1 or higher.

Figure 1 displays a screenshot of the folk ontology opened in Protégé 4.1.

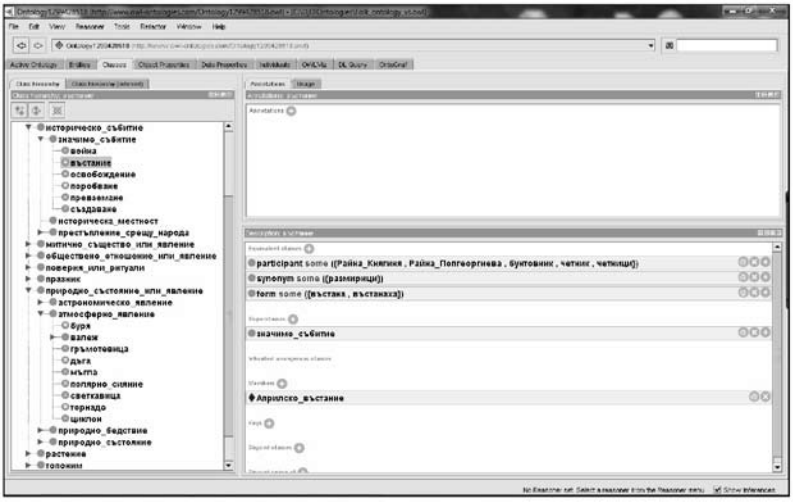


Figure 1. Screenshot of our folk ontology opened in protégé 4.1

3 How the search engine works

In this section we present how our search engine works, what objects are used and how they communicate each other; what information is transferred between them and how search results are posted back to the user.

At the phase of our application development, when the engine is started two objects are created - OntologySercher object and XMLSearcher one.

The OntologySearcher object holds the ontology file, which is loaded into it during the program start; classes, their subclasses and members which will be offered to the user to search in folk lyrics for. The XMLSearcher object will hold information about songs that match the search criteria and data about wrong XML files, if any.

In case of successful start the application publishes a list of available search topics and the user picks one of them. If the program cannot find the ontology file, an error message and short instructions, how to solve the problem, are shown on the page.

After user's click on the chosen topic, all classes, related to this topic, are taken from the ontology, loaded in ASP.NET Treeview control and posted on the page. At this step we invoke two methods of OntologySearcher object consecutively: SetClasses(String root), which gets classes from the ontology and sets them into object's ontologyClasses property (here root is the zero-leveled class related with the chosen topic); and GetClasses() – returns the found classes, i.e. ontology-Classes property.

Next steps are to choose classes from the treeview control and to click confirmation button ("Потвърждавам избраните критерии"). What happens on the page? The treeview control goes in disabled mode (see Figure 2); the confirmation button disappears; a button for enabling the treeview control is shown ("Разреши избор от дървото"); one more button is rendered on the page - the begin search button ("Започни търсене"); members of the chosen classes, if any, are posted on the page, so user can search for them, too. What happens in the code behind? SetClassMeme bers(List<string> classes) method of the OntologySearcher object is invoked. It gets from the ontology all members of the given classes. We invoke and GetClassMembers() method of the same object, too. It returns taken classes' members

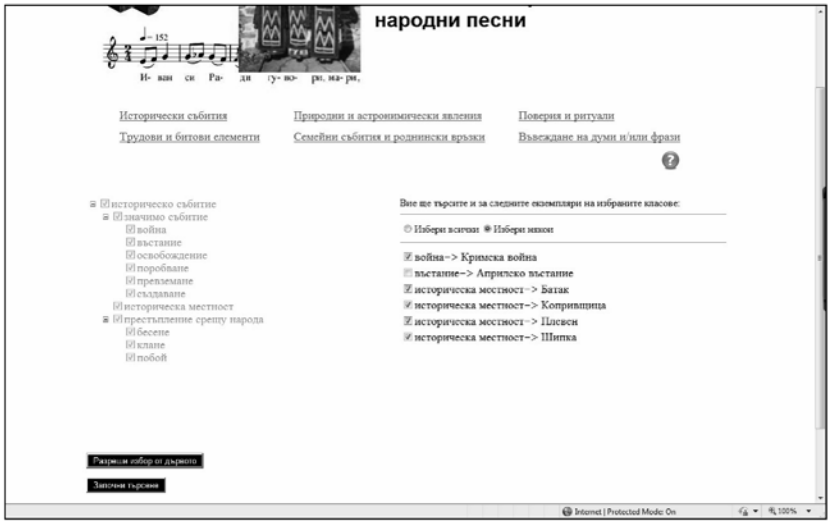


Figure 2. A screenshot just before the user performs a search

Now, let's perform some search by clicking the search button ("Започни търсене").

Here we invoke SetEquivalentClasses(List<string> words) and GetEquivalentClasses() methods of OntologySearcher object. The first method retrieves the equivalent classes of the given ones from the ontology. The second one returns the equivalent classes.

At this step for first time the XMLSearcher object is used. Its methods SetFoundXMLs_searchForWordInText(List<string> words, bool allInLyric) and GetFoundXMLs() are invoked. The first method looks for given words (classes that do not have subclasses) in lyrics and for each lyric/text that matches the search criteria saves its title and found words. The allInLyric parameter says whether we want all words to be in the text or not. We get the found XML files with the second method and show them in an appropriate way to the user.

If some XML file cannot be parsed, the information about the error is stored in a special property named wrongXMLs. We get them, if any, and show to the user in an appropriate way.

On the page we publish the user's choice of search words. There is a condition to enter one word in the search words list - the word to own at least one equivalent class.

When the user clicks on a song's title, on the right can be seen: the path to the XML file describing the song; the song's title and text. Found words in the text are marked in yellow color (see e.g. Figure 3). One more button ("Покажи XML файла с метаданните и текста на избраната песен") is rendered at that moment

- it shows the XML's content in a new window, as displayed on Figure 4. The number of found songs is published, too.

Figure 3. Search words, results and one song's text

Figure 4. The content of selected song's XML file

Figure 5 shows the sequence diagram of the search process.

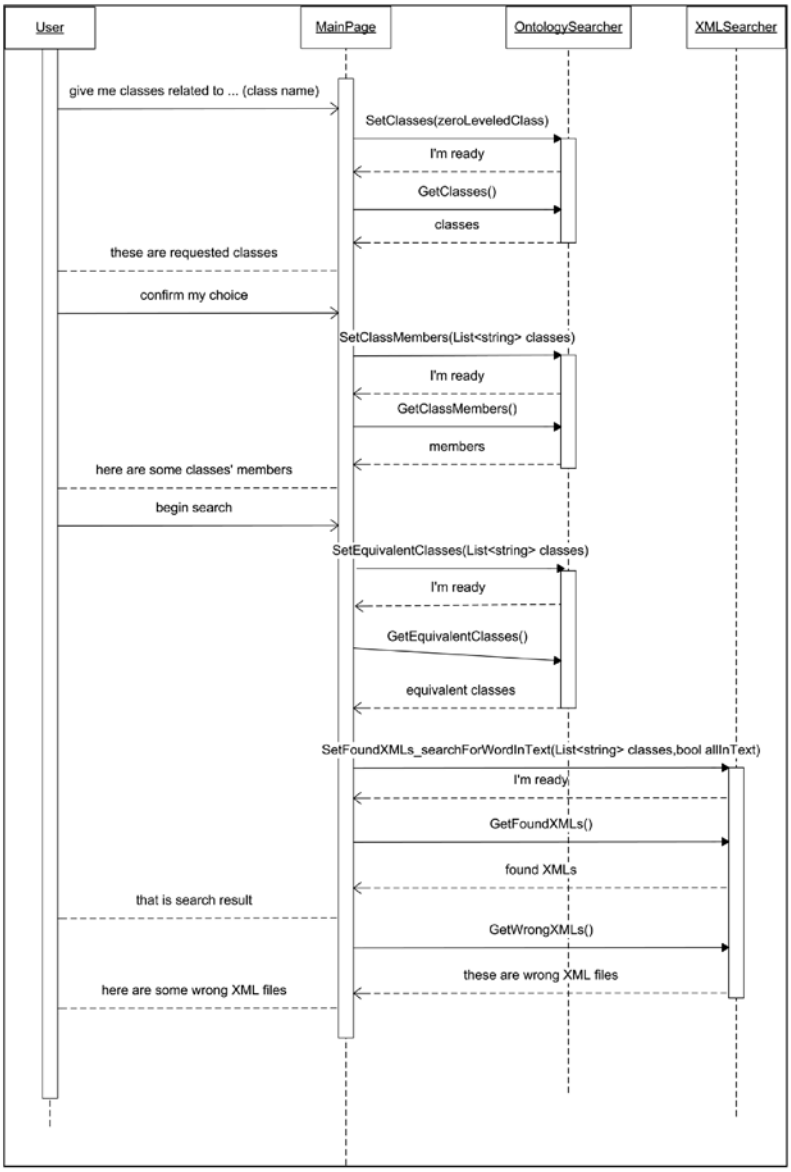


Figure 5. Sequence diagram of the search process

4 Conclusion

As a conclusion we shall mention some directions of our future activities within the discussed project: the search engine's functionality will be extended

with searching in the metadata of songs, its user interface will become more attractive and the engine's performance will be improved.

References

1. Peycheva, L., N. Kirov, M. Nisheva-Pavlova. Information Technologies for Presentation of Bulgarian Folk Songs with Music, Notes and Text in a Digital Library. Proceedings of the Fourth International Conference on Information Systems and Grid Technologies (Sofia, May 28-29, 2010), St. Kliment Ohridski Uiversity Press, 2010, ISBN 978-954-07-3168-1, pp. 218-224.
2. The official ASP.NET site maintained by Microsoft, http://www.asp.net/
3. Nisheva-Pavlova, M., P. Pavlov. Semantic Search in a Digital Library with Bulgarian Folk Songs. In: Y. Tonta et al. (Eds.), "Digital Publishing and Mobile Technologies. 15th International Conference on Electronic Publishing, June 22-24, 2011, Istanbul, Turkey", Hacettepe University, Ankara, 2011, ISBN 978-975-491-320-0, pp. 103-109.
4. Horridge, M., S. Brandt. A Practical Guide to Building OWL Ontologies Using Protйgй 4 and CO-ODE Tools, Edition 1.3. University of Manchester, 2011.

Influence of Different Knowledge Management Techniques and Technology on Organization Effectiveness

Natasha Blazeska-Tabakovska[1],Violeta Manevska[1]

[1]Faculty of administration and management information systems
University "St. KlimentOhridski"
Bitola, Republic of Macedonia

Abstract. Organizations consider the knowledge as a very important resource in recent years, so they have been using a different information technologies and techniques in aim to improve their knowledge management processes in terms of effectiveness and efficiency. This results with a competitive advantage. Some information technologies and techniques don't lead to the preferred effects. This paper analyses the influence of the model on the one organization effectiveness criteria, the improvement of the client satisfaction**.**

Keywords: knowledge management, information technologies, information techniques, organization effectiveness, client satisfaction.

1 Introduction

The new information age has caused drastic changes in the way of manage the organizations and their processes compare to the industrial era. Today, knowledge is the most important organizational resource and its maximum utilization is a key element for any organization that strives to work effectively, it requires knowledge management. Knowledge management is focused on the acquisition of new knowledge from external sources, generating new knowledge within the organization, standardization of existing knowledge in the form of procedures, protocols, transforming individual knowledge into collective process and facilitate reuse and incorporation into processes . These processes can be supported by a variety of informatics technologies and techniques. Different informatics techniques and technologies applied in the function of knowledge management give different effects on improving organizational effectiveness.

We proposed a model which includes: the use of data bases of good practices and lessons learned; the use of data mining for extraction of knowledge about customer with aim to improve the productions / services; extraction of knowledge about customer for marketing purposes; discussion forums and knowledge maps and examine its impact on increasing customer satisfaction.

2 Method of Testing the Model

This model is testing by using a multiple regression. By help of this statistical method it was determined the interconnectedness between increasing customer satisfaction and model.Itwas determinedhow well the model (specific set of variables: the use of data bases of good practices and lessons learned; the use of data mining for extraction of knowledge about customer with aim to improve the productions/services; extraction of knowledge about customer for marketing purposes; discussion forums and knowledge maps)can predict the outcome and which of the variables of the model is the best predictor. It was used a standard regression that means all independent variables were imported at once in the model and individually was measured predicative power of each variable, i.e. how each of the variables, by its input in the model, individuallywill improve it.

The number of independent variables whose influence was explored on increasing customer satisfaction, was determined according to the formula N>50+8*m[1], which gives the relationship between sample size (N) and the number of independent variables (m). The number ofdifferent informatics techniques and technologies (independent variables) that was tested was five, because of analyzed sample size[2],

At first it was determined whether the preconditions for the application of this statistical methodare accomplished. It means to check the relationship between the independent variable, to check whether variables are not multicolinear, among them is no strong correlation[3] and it will be checkednon-singularity of the variables[4], Also it will be checkedwhether there are atypical points; normality and homogeneity of variance.Once determine fulfillment of the requirements goes on the further analysis of the data.

3. Interpretation of the Research Results in Terms of Knowledge Management in Macedonian's Organizations

The research was conducted on sample of 103 organizations in Republic of Macedonia under the type of activity: manufacturing (27,2%), services (65%) and commercial (7,8%); under the ownership: private (73,8%) and public (26,2%); under geographic position: national (57,3%), multinational (33%) and global (19,7%). The research include those who were directly involved in the process of

---

[1] Tabachnick, Fidell(2007, str.123)

[2] The research was conducted on sample of 103 organizations in Republic of Macedoniafrom the private and public sector; **103**>50+8***5**

[3] Strong correlation be considered if r> 0,9

[4] The independent variable can not be represented as a combination of the other variables

analysis, planning, development and implementation of knowledge management system as well as those who feel the effects of the implementing changes.

Although knowledge management is a relatively new science, 80% of respondents were familiar with the program knowledge managementthat facilitating answering the questionnaire. But research has shown that there is a unified name for this program, 70% of respondents this program recognized under the name of managing intellectual capital, 23% in this program recognized learning organization, 2% recognized patent management, while 5% recognized TQM.

In terms of identifying where the capital of knowledge is situated in Macedonian organizations,this researchshow that 53% of organizational knowledge is embedded in products and services. This finding indicates the need for greater use of organizational knowledge. The employees have ideas, possess competencies and skills more than the organization utilizes. Only 22% of organizations knowledge held by employees simultaneously is stored on paper, in computer and incorporated into processes and services.

Knowledge, as a key part of quality tasksperformance, is applied in all business and functional areas of organizations, but not equally. Special emphasis on the use of knowledge is put into marketing and sales sector where particular trade organizations have a high percentage of the use of knowledge; even 87.5% of thesurveyed organizations intensively use knowledge. In the three types of activities, the knowledge commonly is used in research and development sector: 73.1% in manufacturing organizations, 56.9% in service organizations and 75.0% in commercial organizations. Also the knowledge is used in administration; in manufacturing organizations is used 61.5%, in service organizations 53.8%, trade organizations 62.5%. The knowledge in human resources sector is mostly used in commercial organizations 75%, while less is used in manufacturing organizations only 50%. Slightest application of knowledge has in the production of goods and services sector and in manufacturing and service organizations. Very little, knowledge is used in manufacturing organizations in the sector of logistics and procurement; only 30% of surveyed organizations areused knowledge in this sector. (Figure 1.)

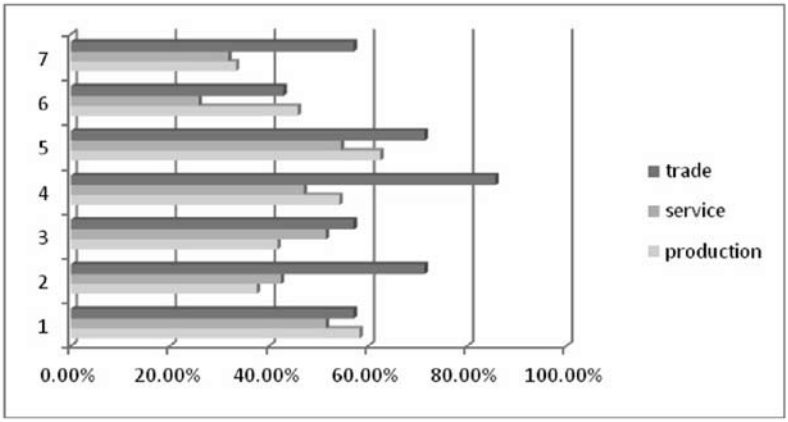


Figure1. Application of knowledge:1-administration; 2-human resource management sector; 3-IT sector; 4-marketing and sales sector; 5-research and development sector; 6-commercial sector; 7-sector of logistics and procurement.

Highest percentage of organizations consider that the biggest benefit of program of knowledge management is increasing the effectiveness (58.8%) and to improving the quality of products / services (43.3%), which is a very high priority for the organizations covered by survey. (Figure2.)

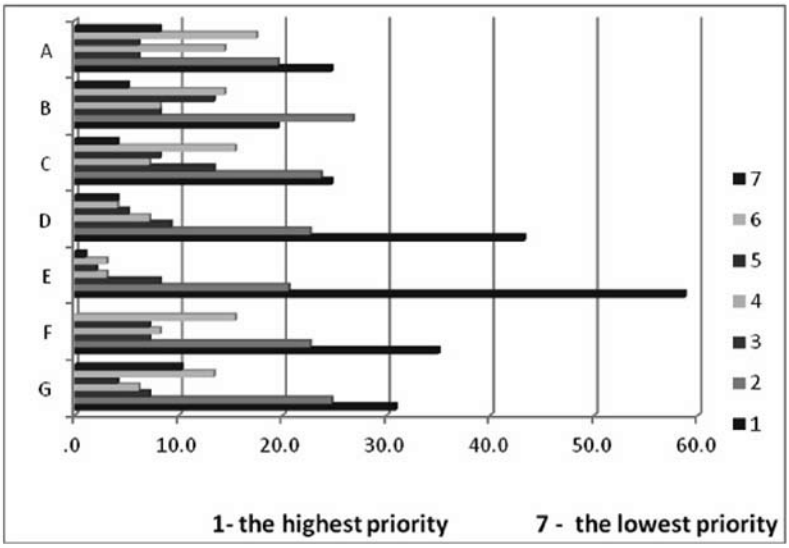


Figure 2. Benefits of Knowledge Management program: A-Improving decision making; B-Decreasing time of delivery; C-Increasing the number of innovations; D-Improving quality quality of products / services; E-Increasing effectiveness; F-Increasing revenue; G-Cost reduction

3.1. Model Valuation.

Table 1. gives a coefficient of determination (r^2 = 0,365)[5], which shows how much of the variance of the dependent variable is explained by the model, i.e.how increasing customer satisfaction can be explained by the model which includes: the use of data bases of good practices and lessons learned; the use of data mining for extraction of knowledge about customer with aim to improve the productions / services; extraction of knowledge about customer for marketing purposes; discussion forums and knowledge maps.

Table 1. Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,604	,365	,332	,347

The tableshows that the coefficient of determination is r^2 = 0,365, which means that our model explains 36.5% of the increase of customer satisfaction. Since this is an optimistic percent, a better estimate gives Adjusted R square = 0,332, indicating that our model supports 33.2% of the increase in customer satisfaction. From the table we can check the statistical significance of this data. The data is statistically significant, becausestatistical significance was p = 0,000[6]

Table 2.ANOVA

ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6,657	5	1,331	11,046	,000
	Residual	11,571	96	,121		
	Total	18,227	101			

3.2. Evaluation of the Individual Impact of the ModelVariables.

The analysis shows how each of the independent variables in the model contributes to the prediction of the dependent variable. It can be seen from the Table 2. that use of data mining for extraction of knowledge about customer with aim to improve the productions / services, individually most contributes to the explanation of the increase customer satisfaction (0440), slightly smaller but

[5] r^2- coefficient of determination, shows how much of the variance of the dependent variable is explained by the model

[6] Thedata is statistically significant if p <0.05

still significant is the use of knowledge maps (0.223). Since p = 0.000 and p = 0.013[7]respectively,we can conclude that both independent variables (extraction of customer knowledge to improve the products / services and the use of knowledge maps) provide a unique and statistically significant contribution on predicting the dependent variable.

Table 3. Values of the coefficients for the independent variables

Model	Unstandardized Coefficients		Standardized Coefficients[8]	t	Sig.	95,0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	,155	,204		,761	,448	-,249	,560
use of data bases of good practices and lessons learned	,044	,081	,048	,542	,589	-,117	,204
use of data mining for extraction of knowledge about customer with aim to improve the productions / services	,398	,087	,440	4,597	,000	,226	,569
extraction of knowledge about customer for marketing purposes	,194	,077	,223	2,536	,013	,042	,346
discussion forums	,073	,091	,071	,804	,424	-,107	,253
knowledge maps	,020	,080	,024	,255	,799	-,139	,179

The previous analysis allows us to answer the posed research question. The model that includes: the use of data bases of good practices and lessons learned; the use of data mining for extraction of knowledge about customer with aim to improve the productions / services; extraction of knowledge about customer for marketing purposes; discussion forums and knowledge maps, supports 33.2% ofincrease customer satisfaction. The largest single contribution from the five independent variables givesextractionof

[7] $p < 0.05$

[8] Standardized coefficients obtained from non-standardizedby converting the values of the variables on the same scale for comparing.

customer knowledge to improve the products / services that individually most contributes to the explanation of the increasingcustomer satisfaction (44%). Slightly smaller but still significant individual contribution provides the use of knowledge maps (22.3%).

4. Conclusion

The model above can be added into existing organizational knowledge management system or can be added in the planning process for implementation of new information system in order to increase organizational effectiveness and increase customer satisfaction. The model includes: the use of data bases of good practices and lessons learned; the use of data mining for extraction of knowledge about customer with aim to improve the productions / services; extraction of knowledge about customer for marketing purposes; discussion forums and knowledge maps

In the analysis of the data for measuring the impact of the proposed model has one of the criteria of organizational effectiveness, and on increasing customer satisfaction, showed that the model explained 30% - 40% of the increase organizational effectiveness, i.e. with their installation as part of existing knowledge management systemor introducing new system, will contribute to increasing the effectiveness.

The greatest impact of customer satisfaction gives customer knowledge extraction for improving products / services (44%) and use of knowledge maps (22.3%).

Based on the above presented results it can be concluded that the practice of customer knowledge to improve the products / services and the application of knowledge maps constitute a major impact on increasing customer satisfaction, one of the criteria of organizational effectiveness.

References

1. Alavi, M., & Leidner, D.: Knowledge Management Systems: Emerging Views and Practicesfrom the Field. Communicationsof the AIS1:5 (1999)
2. Andreu, R., & Ciborra, C.: Organizational Learning and Core Capabilities Development: The Role of Information Technology. Jurnal of Strategic Information Systems (1998)
3. Cooper, D.R.,&Schindler, P.S.: Business research methods. 8th edn. Boston: McGraw-Hill (2003)
4. Liebowitz, J.: Building organizational intelligence: A knowledge management primer. CRC Press LLC (2000)
5. Nonaka, I,: Dynamic Theory of Organizational Knowledge Creation. Organization Science(5:1). (1994)
6. Pallant, J.: SPSS Survival manual: A step by step Guide to Data Analyzes 3th edn. Alen&Unwin (2007)
7. Pentland, B. T.: Information Systems and Organizational Learning: The Social Epistemology of Organizational Knowledge Systems. Accounting, Management and InformationTechnologies

IT Aspects of the Application of Ontologies in Academic Digital Libraries

Daniela Kjurchievska

Technical Faculty, University of Bitola „St. Kliment Ohridski“
daniela.curcievska@ulko.edu.mk

Abstract. The extensive development of digital libraries (DLs) over the last two decades is hardly surprising. Their uses significantly advance the speed of information access. In this point, the demand for digitization of paper based information into digital format is evolving. The recent development of DLs was based on the capability to hold and store a huge amount of digital data. Today we are facing with the challenge to resolve many problems such as interoperability due to heterogeneous data, knowledge for information discovery and partial automation. In such systems ontologies play a major role to cope with these problems.
Within the context of academic digital libraries (ADLs), ontologies can be used to: (i) organizate bibliographic description (bibliographic ontology), (ii) represent and expose document contents (ontologies for context structures), (iii) share knowledge amongst users (community-aware ontologies).
Otherwise, many authors suggest apply ontologies in knowledge management systems in order to improve information search and retrieval and in the same time transform any standard ADL into Semantic Academic DL.
When we are talking about ADLs, it is important to mention the requisites of personalization systems which use all information relevant to the process of searching and browsing an ADL to build a complete navigation profile for each user and its semantic description by means of ontology. Then all those profiles are combined with the help of an ontology that establishes the possible relationships between all the elements present in some future scenario of use in ADL integrated in an e-learning environment.
In this paper are present and discuss some application areas where ontologies have successfully been used in existing semantic digital library systems, but the same applications could be used in building of academic DL.

Keywords: digital library, ontology, semantic technology

1 Introduction

The understanding of a digital library (DL) differs depending on its specific users. A DL is a Web-based electronic storage and access environment for information stored in the digital format, either locally in the library, in a group of networked libraries, or at a remote location. It also represents an integrated set of services for capturing, cataloging, storing, searching, protecting and retrieving information. It comprises digital collections, services, and infrastructure to support lifelong learning, research, scholarly communications, and preservation.

The common opinion is that the Web is not a digital library because digital libraries are perceived as libraries with the same purposes, functions, and goals as traditional libraries, that is, collection development and management, subject analysis, index creation, provision of access, reference work, and preservation. Due to its inherent complexity, the current tendency in building DLs is to move forward in small, manageable and evolutionary steps, rather than in a rapid revolutionary manner.

Nowadays, we are finding new types of libraries coming up from long-term personal libraries, as well as DLs that serve specific organizations, educations needs, and cultural heritage and that vary in their reliability, authority and quality. Besides, the collections are becoming more heterogeneous in terms of their creators, content, media and communities served. In addition, the user communities are becoming heterogeneous in terms of their interest, backgrounds, and skill levels, ranging from novices to experts in specific subject areas [1]. This growing diversity has changed the initial focus of providing access to digital content and transforming the traditional services into digital ones to face the problem, whereas the next generation of libraries should be more proactive offering personalized information to their users taking in consideration each person individually (his or her goals, interests, level of education, etc.).

While data and information are captured and represented in various digital formats, and rapidly proliferating, the techniques for accessing data and information are rudimentary and imprecise, mostly based on simple keyword indexes and relational queries. In the current context of explosive availability of data, there is a need for knowledge discovery approach, based on both top-down knowledge creation (e.g. ontologies, subject heading, user modeling) and bottom-up automated knowledge extraction (e.g., data meaning, text meaning, web meaning). It promises to help transfer DL from institution of data and information to an institution of knowledge [2].

In that sense, building an Academic Digital Libraries (ADLs) is very important for the academic community. Justifications for the development of ADLs include the desire of preserve science data and the promises of information interconnectedness, correlative science, knowledge for information discovery, and system interoperability [3], [4]. Application of ontologies in ADLs is fundamental to fulfilling those promises. This paper will present and discuss some of the application areas where ontologies have successfully been used in existing semantic DL systems.

2 Academic Digital Libraries

In present year, many higher institutions provide academic digital libraries. Kalinichenko et al. [5] noted that libraries may transform the way we learn, providing supporting resources and services, operating as decentralized but integrated/

virtual learning environments that are adaptable to new technologies. So, ADLs are those libraries that serve the information needs of students and faculty of the colleges and universities. By definition, ADLs play a very crucial role in bridging students, academicians and researchers needs of information in this borderless information seeking era. Academic and intellectual endeavors may be supported by ADLs towards, not only simply for information seeking, but also for exploring, researching and enlarging their knowledge via adapting the information systems and human-computer-interacting technologies.

3 Applications of Ontologies in Digital Libraries

3.1 Using the Annotation Ontology in Semantic Digital Libraries

Semantic Digital Libraries (SDLs) make extensive use of meta-data in order to support information retrieval and classification tasks. Within the context of SDLs, ontologies can be used to (i) organize bibliographic descriptions, (ii) represent and expose document contents, and (iii) share knowledge amongst users [6], [7].

Concerning the organization of bibliographic descriptions, Kruk et al. [6] proposed to lift bibliographic metadata to a machine-interpretable semantic level by applying concept defined in an ontology, which in case of JeromeDL [8] allows users to semantically annotate books, papers, and resources. In Bricks [9], they proposed a different approach for bibliographic ontologies in order to support arbitrary metadata formats and enable management of metadata that describes contents in various, domain specific ways.

Modern digital library systems not only store bibliographic metadata but also store electronic representation of the content itself. In order to represent and expose document content, a universal layer for metadata and content retrieval was provided, by including structural concept in ontologies and using those concepts in metadata descriptions. This approach enables an easy extension of structure description, of resources with new concepts, without changing the underlying database schema or violating the integrity of existing data.

In the context of shared knowledge amongst users, a community-aware ontology has been proposed. In this approach the main goal is to share knowledge within groups of users, so that each user can utilize and learn from the experience of other users. This ontology contains a unified way for describing users, allowing to specify friendship relations among properties like name, location or interests, and offers sharing bookmarks and catalogs between friends, thus provides a base for social semantic collaborative filtering. This proposed ontology should serve as a mechanism to describe the knowledge of users and communities so that DLs can perform the step from static information to dynamic knowledge spaces.

3.2 Ontologies and Multilingual Academic Digital Libraries

Another case is where ontologies play a significant role in the implementation of the full functionality of semantic oriented search engine in multilingual academic digital library. A successful example is developing of DigLib-CI, a digital library created at the Department of Computer Informatics of the Faculty of Mathematics and Informatics at the Sofia University [10], [11]. There are two subject ontologies included of DigLib-CI, the Computer Science ontology and the Information Systems ontology, based on the Computer Sciences Curriculum 2008 of ACM and IEEE [12] and the Model Curriculum and Guideline of Undergraduate Degree Programs in Information Systems of ACM, AIS and AITP [13] respectively. They provide the development of an adequate search engine with complete viewpoint towards the conceptual structure of areas of Computer Sciences and Information Systems. All the resource descriptions consist of two equivalent parts in which the element value are text in Bulgarian and English respectively, so the search engine examines the corresponding part of description to the language of the user's query. So this proposed approach provided facility for flexible semantic-oriented access to the library resources for users with various professional profiles and language skills.

3.3 Personalization System in Digital Libraries

A different aspect of application of ontology in DLs is the description of a browsing and searching personalization systems. It is based on the use of ontologies for describing the relationship between all the elements which take part in digital library scenario of use. In this case, it is important to clarify that ontology is not used for describing the contents of a digital library, but for describing the way users browse and search such contents, with the aim to build a personalization system based of accurate recommendations. Personalization is one of the key factors which are directly related to user satisfaction [14] and, therefore, linked to the failure or success of the performed activity, although it must be carefully introduced Ferran et al [15] describe the set of desired functionality and requirement of scenario for a digital library which includes personalization capabilities by means of ontologies. They used ontologies for describing the possible scenarios of use in DL, bringing the possibility of predicting user requirements in advance and to offer personalized services ahead of express need. The elements that determine the functionalities of this personalization system are the user's profile, which includes navigational history and its preferences, and the information collected from navigational behavior of the digital library users. Beside those, they also identified other basic elements such as: navigational profiles, user actions and the relationships between these elements as a part of the ontology which is used by the personalization system.

3.4 Knowledge Management System

Marjit et al. [16] suggest the framework of ontology-based Knowledge Management System in order to improve information search and retrieval. This concept is applicable for any existing DL to make the transition from DL to semantic DL. Semantic DLs offer expanded facilities for knowledge discovery, data mining of semi structured text, and mechanisms for linking and searching related concepts. The main objective of proposed architecture of OKMS is as follows:

- Storage of digital object within a Web Server to prepare the DL,
- Classification of documents (DCU),
- Preparation of domain ontology to describe the working domain with semantics (ODU),
- Making the DL more machine and user friendly (GUI, Interactive Graphical interface)
- Development of ontologies to describe each digital object semantically,
- Better information discovery.

Fig. 1 describes the conceptual framework which was proposed to introduce the ontology based knowledge management in the existing digital libraries in order to transform any standard digital library into SDL.

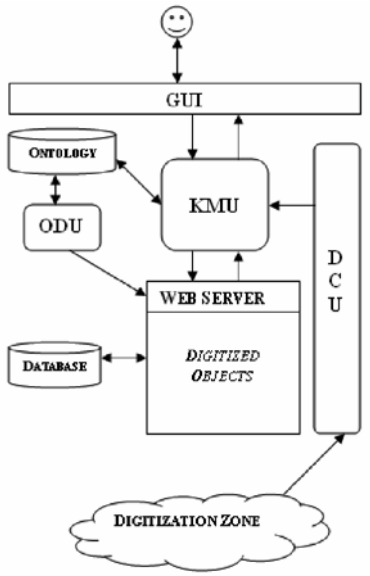


Fig. 1. Conceptual model of framework for ontology based Knowledge System for SDL.

A dedicated knowledge management unit is proposed for the efficient and effective knowledge management within the DL. It also handles the user requests for digital documents or information in digital format. But the main job of this

module is to perform search operations to find out and retrieve the accurate information both on the user's requirements based on semantic matching between user request and on ontological description of working DLs and their contents.

4 Conclusion

An academic digital library plays a very crucial role in bridging students, academicians and researchers' needs on information in this borderless information seeking era. So justifications for the development of ADLs include the desire to preserve science data and the promises of information interconnectedness, correlative science, knowledge for information discovery, and system interoperability. Application of ontologies in ADLs is fundamental in fulfilling those promises.

In this paper, some of the application areas for using ontologies in semantic digital libraries were presented as follows:

- Bibliographic ontology
- Ontology for context structures
- Community-aware ontology
- Ontology based searching tool in multilingual ADL
- Personalization system for DL
- Ontology-based Knowledge Management System

All those application areas where ontologies have successfully been used in existing DL systems should be taken into consideration in building an ADL in order to improve its performances and efficiently and effectively supporting academic and educational tasks.

References

1. Callan, J., Smeaton, A.: Personalization and Recommender Systems in Digital Libraries. Joint NSF-EU DELOS Working Group Report, available at: www.dli2.nsf.gov/international projects/working_group_raports/personalisation.html. (2003)
2. Chen, H.: Towards building digital library as an institution of knowledge. NSF Post Digital Library Futures Workshop, Chatham, MA, available at: www.sis.pitt.edu/%7Edlwkshop/paper_chen.html (2003)
3. Hughes, S.J., Crichton, D.J., Mattmann, C.A: Scientific digital libraries, interoperability, and ontologies. Proceedings of the 2009 Joint International Conference on Digital Libraries, JCDL 2009, June 15-19, 2009, Austin, TX, USA (2009)
4. Hughes, S.J., Crichton, D.J., Mattmann, C.A: Ontology-Based Information Model Development for Science Information Reuse and Integation . Proceedings of the IEEE International Conference on Information Reuse and Integration, IRI 2009, 79-84 (2009)
5. Kalinichenko, L.: Dgital Libraries in Education: Analytical Survey. Moscow, UNESCO Institute for Information Technologies in Education, 2003
6. Kruk, S., Haslhofer, B., Piotr, P., Westerski, A., Woroniecki, T.: The Role of Ontologies in Semantic Digital Libraries. European Network Knowledge Organization Systems (NKOS) Workshop, Spain (2006)
7. Garcia-Castro, L.J., Giraldo, O.,X, Castro, A.G.: Using the Annotation Ontology in

Semantic Digital Libraries. Proceedings of the ISWC 2010 Posters & Demonstrations Track: Collected Abstracts, Shanghai, China, November 9, 2010
8. Kruk, S., Decker, S., Zieborak, L.: JeromeDL – Adding Semantic Web Technologies to Digital Libraries. 2009
9. BRICKS Project: Building Resources for Integrated Cultural Knowledge Services (IST 507457), http:/www.brickscommunity.org
10. Nisheva-Pavlova, M., P. Pavlov. Search Engine in a Class of Academic Digital Libraries. In: T. Hedlund, Y. Tonta (Eds.), "Publishing in the Networked World: Transforming the Nature of Communication. 14th International Conference on Electronic Publishing, 16-18 June 2010, Helsinki, Finland", Edita Prima Ltd, Helsinki, 2010, ISBN 978-952-232-085-8, pp. 45-56. (2010)
11. Nisheva-Pavlova, M. Providing and Maintaining Interoperability in Digital Library Systems. Proceedings of the Fourth International Conference on Information Systems and Grid Technologies (Sofia, May 28-29, 2010), St. Kliment Ohridski Uiversity Press, 2010, ISBN 978-954-07-3168-1, pp. 200-208. (2010)
12. ACM, IEEE Computer Society. Computer Science Curriculum 2008 (An Interim Revision of CS 2001), http://www.acm.org/education/curricula/ComputerScience2008.pdf, last accessed on August 16, 2012.
13. ACM, AIS, AITP. IS 2002 (Model Curriculum and Guidelines for Undergraduate Degree Programs in In-formation Systems), http://www.acm.org/education/is2002.pdf, last accessed on August 16, 2012.
14. Riecken, D.: Personalized view of personalization. Communications of AMC, Vol.43 No.8, pp.27- 28. (2000)
15. Ferran, N., Mor, E., Minguillon, J.: Towards personalization in digital libraries through ontology. In Library Management. Vol.26 No. 4/5, 2005, pp. 206-217 (2005)
16. Marjit, U., Sarkar, A.: Ontology based Management System and Its Application to Semantic Digital Library Initiative. 8th International CALIBER-2011, Goa University, Goa, Mrch 02-04, 2011

A Method for Decision Making in Computer Science Research

Neli Maneva
Institute of Mathematics and Informatics, BAS,
Acad. G. Bonchev Str., Bl.8, 1113 Sofia, Bulgaria,
neman@math.bas.bg

Abstract. The paper describes the peculiarities of the project-based scientific research and shows how the method of Comparative analysis (CA) can be used as a decision support method during such research. The CA method and the procedure for its application are briefly described. Taking into account the specific features of scientific research, the relevant types of compared objects (products, processes and resources) have been identified. Some results of the CA use in a real-life scientific project are presented, confirming the feasibility of the approach. Finally, a summary of the obtained results is given and some ideas how the specified approach can be further developed, are mentioned.

Keywords: Project-based scientific research, multi-criteria decision making, comparative analysis, business rules extraction.

1 Introduction

Computers and computerized devices are unavoidable part of professional and private life of millions people all over the world. Most of these devices are sophisticated software-intensive systems and their production is accompanied with many problems - a great number of postponed and over-budgeted projects, developing software products, which often don't meet entirely the expectations of their users. This situation introduces some new requirements to the process and results, obtained from the scientific research in the field of computer science. Instead of traditional research without rigid financial and time constraints, now many research activities have to be accomplished within projects with clearly stated goals, budget and duration, pre-defined by the funding sources. The new project-driven research changes entirely the style of scientific work, especially in the field of computer science, where the pressure for innovative and transferable to practice results is high. In order to provide the desired project's deliverables in time, the scientists, involved in such projects, should enrich their professional competencies with a piece of management knowledge and some special techniques and skills, facilitating the work on the project. One such skill, significant from

pragmatic point of view, is the ability to make a prompt and reasonable, multi-criteria choice in any decision making situation.

In this paper we describe our attempt to bring together a rigorous decision support method and scientific research. Next section presents the basic characteristics of the project-based scientific research. Section 3 describes the Comparative analysis method and how it can be used for scientific research, shaped in projects. Section 4 summarizes the results of experimental use of the proposed approach in a real-life scientific project. In Conclusions some ideas for author's intentions in this direction have been mentioned.

2 Project-based Scientific Research

The efforts to make science more meaningful and successful have been recognized as important both at European and national levels [6]. In search for excellence in performance, some new funding schemes and managerial approaches to scientific research have been promoted thus changing entirely the style of scientific work. Instead of traditional "blue skies" research, defined as "research without a clear goal" or "curiosity-driven science" [1], many scientific organizations now should adapt the project-based scientific research [7]. Summarizing the ideas of [5], we can describe its main characteristics as follows:

- Scientific project represents some temporary efforts, undertaken to create a unique product or service, or to introduce innovation;
- Research process, shaped in a project, is unique and comprises a set of coordinated and controlled activities performed so as to achieve the preliminary defined goals;
- Project implementation is under triple limitations - in project scope, time and cost;
- Project is carried out in conditions of uncertainty and should involve the risk analysis and management activity.

The effective and efficient achievement of project objectives within the defined constraints requires significant managerial efforts both at a strategic level - doing the right things, and at an operational level - doing things right. For scientists, working on projects, some additional managerial competencies will be helpful, e.g. to be acquainted at least with the project management framework, defined in ISO 10006:1997 [3] and describing the basic knowledge, skills, special tools and techniques, applied to project activities. This standard collection can be extended with other items, proved to be useful and in the next section we will present a such one.

3 Comparative Analysis in Project-based Scientific Research

We are going to show that a formal and systematic approach can improve the

decision-making so as to achieve the most significant from managerial point of view objectives: to decrease the costs, to improve the team performance, and to shorten the duration of the scientific project.

3.1 Comparative Analysis – a Brief Description

Let us introduce a formal method for a reasonable choice. It supports the multi-criteria decision making and has been developed in accordance with the best achievements in this area [3].

Comparative Analysis (CA) is a study of the quality content of a set of homogeneous objects and their mutual comparison so as to select the best, to rank them or to classify each object to one of the predefined quality categories.

The compared objects can be products, processes or resources, identified as significant for the activity under consideration. When apply the CA method, we distinguish two main players: the ***Analyst***, responsible for all methodological and technical details of CA implementation, and a ***CA customer*** - a single person or a group of persons, who should make a decision in a given situation and wants to use the CA as a supportive tool. The context of the desired CA is specified through a ***case***, described by the following six elements:

case = { View, Goal, Object, Competitors, Task, Level }

The **View** describes the customer's role and the perspective from which the comparative analysis will be performed. For project-based scientific research the customers are with different background, qualification and status as a decision makers in the institutions, involved in the project – the funding one, the organization, responsible for project accomplishment, members of the research team, etc.

The **Goal** expresses the main customer's intentions in CA use and can be to describe, analyze, estimate, improve, predict or any other, formulated by the Customer, defining the case.

The **Object** represents the item under consideration. Especially for project-based scientific research in computer science the analyzed objects can be:

- Products - different intermediate or final project's deliverables;
- Processes, related to the whole project life cycle - from appearance of idea to be investigated till the project follow-up analysis;
- Resources: project-oriented, technological, etc.

For each investigated object a quality model should be created – a set of characteristics, selected to represent the quality content, and the relationships among them. This modeling activity is the most difficult due to its high cognitive complexity.

According to the goal, the set C of **Competitors** – the instances of the objects

to be compared – should be chosen. Usually the stated goal is to perform the Comparative analysis so as to obtain the ranking of a number of objects and then the set C comprises those competitive objects.

The **Task**, described as an element of a case can be Selection (finding the best), Ranking (producing an ordered list), Classification (splitting the objects to a few preliminary defined quality groups) or any combination of them.

The depth **Level** defines the overall complexity of the CA and depends on the importance of the problem under consideration and on the resources planned for CA implementation.

3.2 A Decision Making through CA in Project-driven Scientific Research

Generally speaking, the CA method can be used in any decision making situation within the project-based scientific research after specifying a case with an appropriate definition of the above mentioned elements. From methodological point of view, first some preliminary steps, specific for the selected activity, should be taken. The Analyst has to identify the main decision makers. In the project-based scientific research they can be at different levels of hierarchy in two organizations (financing and accomplishing the project) or representatives of stakeholders - customers or clients, for which the project is carried out. The following roles of CA Customer have been identified till now:

- Contractor, observer, reviewer (evaluator) and contact person - representatives of the funding organization;
- CEO, project manager and scientific policy maker - from the upper management of the organization, which implements the project;
- Project leader, adviser, researcher, member of the technical staff – from project's team;
- Stakeholders – institutions or individuals, who are going to use the results of the project.

Taking into account the responsibilities and typical tasks of these players, a number of cases can be defined, reflecting their point of view to the analyzed situation.

Next during the CA preparation stage, a few basic objects should be identified and their respective hierarchical quality models should be constructed. For project-based scientific research in the field of computer science the following basic objects, classified in three groups, have been identified:

Products. We consider the typical deliverables in any scientific project (surveys, papers, reports, models, hypothesis, case studies, experiments, and prototypes) or some deliverables, specific for CS projects like requirements, specifications, designs, software components and systems, documentation, developer's stories, etc. The detailed description of the created models is beyond the scope of this

paper, but from pragmatic point of view it is important to mention that all created models are saved in a CA repository so as to facilitate their re-use.

Processes. Modifying slightly the process classification, proposed in [5], we consider five main groups of processes within the project, for which the CA seems to be fruitful as a method, supporting the decision making:

- Initialization process, during which the decision how to execute the project should be made;
- Planning process, defining the goals, the criteria for project success and clarifying the alternatives for achieving the goals;
- Implementation process, involving coordination of people and other resources so as to achieve the stated goals effectively;
- Control process, covering regular analysis of the plan and deciding on the necessity of corrective measures, their determination, coordination, fulfillment and assessment;
- Project closure process, identifying the formal procedures for completing the project and submission of its final deliverables.

For the purposes of the CA, a basic process model has been constructed. It is an abstract representation of a process, presenting its quality content through a hierarchical structure over a set of process quality characteristics. A number of derivative models, reflecting the peculiarities of the above mentioned five processes have been created, taking into account the CA context, described by a particular case.

Resources. We consider two groups of resources: typical for any project (money, people, physical environment) and technological (approaches, methods, techniques, tools). In scientific projects, in which the success depends highly on applied methodology, modeling of the objects from the second group is crucial for the CA use, thus assuring the proper choice of the most appropriate research methods and tools.

4 A Case Study: CA in a Real-life Scientific Project

The feasibility of the approach, described above, has been examined within a scientific project "Automatic Business Rules Extraction from Programs", performed under a contract with the National Scientific Research Fund.

Following the methodology for application of the CA method, described in [3], in order to adopt it for project-based research, the content (as a set of mutually connected activities) of three main stages should be defined. During the ***Preparation*** stage we have to define a problem, encountered in Business Rules (BR) extraction project. The Analyst communicates intensively with the Customer so as to clarify the situation and to describe precisely a sequence of cases to support the decision making. Technically, for each case the Analyst

checks which case elements have already existed from previous analysis and can be subject only for modification, and which have to be constructed from scratch. In this stage the involvement of the decision maker, acting as a CA Customer, is highly recommended. During the **Implementation** stage the Analyst reuses or constructs the needed object models, forms the set of Competitors, and performs the CA Task, using the appropriate software tools. At the **Follow-up** stage the obtained results have been analyzed so as to create a concrete action plan to solve the problem under consideration.

The usefulness of the CA as a decision support method will be illustrated by three problems, encountered during the project.

Problem 1. How to define the most appropriate work position (Role) for members of the group, involved in BR extraction?

First, on the basis of the Job description, a profile has been created for each of the identified Roles: Business Analyst, Policy Manager, Software Architect, Software Developer, System Administrator, Policy Translator, Rules Extractor and Mediator. The profile comprises three groups of characteristics: professional competences, soft skills and experience. Then the CA has been used to produce an ordered list of candidates for each Role through comparison between the role profile and the individual profiles of team members.

Problem 2. How to create a short list of research papers, covering some topics, investigated within the scientific project?

This problem is significant, especially at the beginning of a scientific project, when the state of the art in the area of research should be clarified. Constructing a list of keywords and using the contemporary search tools, usually a huge amount of papers are identified as potential sources of information. Due to preliminary defined and insufficient project resources it is quite natural only part of them to be selected for further study and review.

The CA has been used for this problem. First, for object "research paper" a simple linear model has been created, comprising five quality characteristics: currency, availability, utility, relevance and validity, evaluated in range [0, 2]. The total quality value (the rank) of a paper is calculated as a weighted sum of characteristics values. Working together, the Analyst and the corresponding CA Customer (project leader, senior research fellow or any member of the research team, responsible for a survey on a given topic) performs the search with an appropriate set of keywords and then evaluate each paper from the obtained list, calculating its rank. The next step is to apply the CA with a generic case

case = { View, Goal, Object, Competitors, Task, Level },

where the values of the elements are as follows:

- **View** – that of the evaluator, who is a member of the research team
- **Goal** – evaluation;
- **Object** – a research paper;

- **Competitors** – the papers from a list, obtained after a search;
- **Task** – ranking;
- **Level** – simple.

As a result an ordered list of all papers has been created, representing the point of view of the evaluator, who can decide about the length of the short list of papers to be studied in detail.

Problem 3. How to select only one method for BR-extraction to be further investigated and examined through prototyping and experiments in a real-life modernization project?

At the milestone, marking the end of the first project phase, the researchers involved in the project, should decide about the main directions of research and experiments during the second phase. For this problem the CA can be used in a sequence of cases with the following values of the main elements:

View: The identified points of view for this problem belong to the following Roles of the involved participants: Business Analyst, Policy Manager, Software Architect, Software Developer, System Administrator, Policy Translator, Rules Extractor and Mediator.

Goal: To assess, to evaluate, to compare

Object: A BR-extraction method

For this object a simple linear model has been created, comprising 4 characteristics: effectiveness, feasibility, efficiency and cost.

Competitors: In all cases the set **C** of Competitors comprises the methods from a short list, identified as candidates for further research and experimental work.

The appropriate **Task** here is the Selection (to find the most appropriate) and the Ranking, so as to produce an ordered list of all compared methods. In all designed cases the **Level** is High, because the results of the performed Comparative analysis will be of crucial significance for the future of the scientific project.

The results obtained till now after CA use within the above mentioned project are encouraging and we are going to identify some other decision making situations in the process of the BR extraction, for which the CA method seems to be fruitful.

5 Conclusions

The purpose of this paper has been to propose a approach to project-based scientific research, unifying a methodology, a procedure for its application and a set of tools so as to support the decision making in situations, identified as significant within a project. The developed approach has been examined in a real-life scientific project thus proving its feasibility and giving the chance for

initial assessment of its advantages and shortcomings. Our main conclusion is that the project-based scientific research is possible, but it should be carried out professionally – in a systematic and tool-supported way so as to guarantee (at least a piece of) success.

Our future research and development activities can be:

- To continue the collaboration with experts in the theory and practice both in scientific research and project management, thus meeting the challenges of the successful CA adoption, especially for constructing models of the relevant objects;
- To identify and describe entirely a number of CA cases for crucial decision points during a project-based scientific research, expanding the CA library of re-used items for this activity – basic and derivative models, metrics, etc.
- To use the approach in other scientific project in order to study its vulnerability to the basic project parameters (scope, cost, duration);
- To investigate how the CA method can be used within software projects, applying modern development approaches, like the integrated one, described in [4], Boehm's scalable spiral model, etc.

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

References

1. Braben, D.: Scientific Freedom: The Elixir of Civilization. John Wiley and Sons Inc. (2008)
2. ISO 10006:1997 Quality management – Guidelines for quality in project management
3. Maneva N.: Comparative Analysis: A Feasible Software Engineering Method, Serdica Journal of Computing, vol.1, N°1, pp.1-12 (2007)
4. Napoli, J.P, Kaloyanova, K.: An Integrated Approach for RUP, EA, SOA and BPM Implementation, Proceedings of the 12th International Conference on Computer Systems and Technologies, Vienna, Austria, 16-17 June 2011, pp.63-68 (2011)
5. Savić, S. et al.: Management of innovation projects in the context of the competitiveness of the organization. Proc. of the International Symposium Engineering Management and Competitiveness (EMC2011), June 24-25, 2011, Zrenjanin, Serbia, pp. 111-116 (2011)
6. From Challenges to Opportunities: Towards a Common Strategic Framework for EU Research and Innovation Funding, May 2011, position paper by the British Academy in response to the consultation on the Green Paper. http://www.britac.ac.uk/intl/european_framework_programmes.cfm.
7. Impact of external project-based research funding on financial management in Universities. Expert Group report, November 2008. http://ec.europa.eu/research/research-eu

DISTRIBUTED SYSTEMS

Digital Libraries and Cloud Computing

Maria M. Nisheva-Pavlova

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

Abstract. Cloud computing is a relatively new information technology which is an improvement of distributed computing, parallel computing, grid computing and distributed databases. The basic principle of Cloud computing is implementing tasks distributed in large numbers of computers and other devices over a network. It allows collection and integration of information and other resources stored in personal computers, mobile phones etc. and putting them on the public cloud for serving users. Recently digital libraries are slowly entering their Cloud computing era. The application of Cloud computing technology in digital libraries will lead to better understanding and more efficient implementation of large part of digital library services. The paper aims to examine the issue of Cloud computing and its contribution to digital library service provisioning.

Keywords: Digital library, Semantic digital library, Cloud computing, Web services.

1 Introduction

During the last 1-2 decades, various universities and colleges that are steadily raising their respective teaching and research level, have established their own digital libraries. These libraries contain learning and research materials like books, papers, theses, and other works which were digitized for the purpose or were "born digital". Academic digital libraries are committed to maintaining valuable collections of scholarly information. Usually information resources between various universities are relatively independent, at the same time inadequate and redundant, and therefore ineffectively used. Because of that, the maximum utilization of digital library resources is a matter of considerable importance.

On the other hand, the subject of Cloud computing has been a hot topic for the past several years. There are many implications regarding Cloud computing for service provisioning for libraries. Several issues involving Cloud computing are explored, and the advantages and disadvantages concerning this technology need to be seriously considered by digital library technology professionals. The paper presents a study of the issue of Cloud computing technology and its implications for providing library services in the near future.

V. Dimitrov (Editor): ISGT'2012. ISSN 1314-4855
Proceedings of the 6th International Conference on
Information Systems and Grid Technologies, Sofia, June 1-3., 2012.

2 Current Issues in Digital Libraries

A number of trends and challenges demonstrate the breadth and depth of research and development in digital libraries. Here we present in brief the most significant part of them.

2.1 Architecture, systems, tools and technologies

Within this category lie all technical, infrastructural, algorithmic and system-related components of digital libraries. Some of the key issues here, according to [1 – 4], are the development and use of:

- open networked architectures for new information environments;
- content management systems;
- novel search and retrieval techniques such as integrating links and ranking;
- audio-visual and multimedia information retrieval systems;
- intelligent systems for indexing, abstracting and information filtering;
- harvesting and interoperability technologies;
- collaborative, visual, 2D and 3D interfaces.

2.2 Digital content and collections

Digital collections require well-structured ***metadata*** schemes to describe digital objects and content at various levels of granularity. Structural and descriptive metadata are two general classes of metadata of particular relevance. One major challenge with regard to metadata is the diversity of digital information formats and the ways in which they should be described in different collections with different target audiences and uses.

Issues for metadata researchers include [1]:

- human and algorithmic approaches to metadata provision;
- choosing from a wide range of metadata formats;
- applying metadata standards across digital collections;
- metadata harvesting;
- mappings between different metadata formats.

Interoperability is one of the most heavily discussed issues in digital library research. Interoperability in general is concerned with the capability of different information systems to communicate. This communication may take various forms such as the transfer, exchange, transformation, mediation, migration or integration of information. The requirement for interoperability derives from the fact that various digital libraries with different architectures, metadata formats and underlying technologies wish or need to effectively interact. The proper application of common protocols, standards and ontologies is indicated as an

effective instrument for providing and maintaining most types of interoperability in digital library systems [5 – 8].

The Open Archives Initiative (OAI) protocol is the most widely discussed and investigated standard for cross-repository interoperability. It allows distributed digital libraries to expose their metadata to a wide range of search and retrieval services and also to extract metadata from Web databases. Z39.50 has also been mentioned as another interoperability protocol for online catalogues and other types of information retrieval systems on the Web.

In addition, a set of other issues related to digital content and collections have been discussed. These include [1 – 3]:

- collection development strategies, policies and management;
- identifying collections of information which are not accessible or usable because of technical barriers;
- formulating strategies for sustainable and scalable collections;
- encouraging the development of new collections;
- the creation of new genres of digital objects;
- issues related to digital preservation and Web archiving.

2.3 Standards

Standards within the context of digital library research encompass all protocols and conventions that have been set for digital library architectures, collections, metadata formats, interoperability etc.

Some types of standards which have been the focus of research include [6]:

- digital collection development standards;
- archiving and preservation standards;
- metadata formats (e.g. Dublin Core, MARC, IMS);
- electronic publishing standards for books, journals and other media.

2.4 Knowledge organization systems

This category refers to a range of tools used for organization, classification and retrieval of knowledge. Digital library researchers operating in different contexts have explored the potential of these tools for different purposes.

Some of the corresponding applications are:

- use of thesauri and classification systems for cross-browsing and cross-searching across various digital collections;
- creation of ontologies using existing thesauri;
- development of classification systems and specialized controlled vocabularies to provide a general knowledge-representation facility for digital collections with a diverse range of materials;
- use of taxonomies to provide unified and organized access to different digital

repositories through describing different layers of the digital collections.

One of the challenges here is the way in which these tools can interact with each other. Research in this direction is on the way to investigate issues concerning mappings and interoperability among various knowledge organization systems.

2.5 Users and usability

In order to develop usable digital libraries and to improve system design, researchers have addressed user behaviour and user requirements in different contexts including academic environments, schools, government departments and business.

The following areas have been the focus of a number of studies [2]:

- empirical studies of users interacting with digital libraries;
- usability, accessibility and user acceptance of digital libraries;
- user-centered support for learning, teaching and research through the convergence of virtual learning environments and digital libraries;
- human-computer interaction;
- evaluation of the behaviour of diverse user communities based on their age, knowledge, and particular needs.

One of the major challenges in user studies is associated with the data gathering methodology and techniques. Researchers have tried to use a combination of tools and techniques to collect useful data for user evaluation.

2.6 Legal, organizational, economic, and social issues

Rights management, intellectual property and copyright issues are the legal aspects of digital libraries. Social issues in relation to digital libraries centre on the ways in which people view digital libraries and their usefulness. There are also economic issues such as commerce, shopping, marketing and business competition, all of which form part of the digital library research discussion. This area touches on topics such as:

- intellectual property in the complex global market;
- legal issues associated with access, licensing, copying and dissemination of digital materials, economic, business and pricing models and strategies;
- sustainability and survivability, new business models and marketing strategies.

Although these are the main research categories, the list discussed so far is not exhaustive. Other issues include:

- evaluation issues out with the users and usability category, reference and question-answering services;
- the development of different types of digital libraries.

Evaluation is a critical issue in digital library research. While user-oriented

evaluation can be discussed in the users and usability category, evaluation also applies to digital library systems, their performance, the underlying technology and the information retrieval techniques utilized.

There are also numbers of projects concerned with the design, development and evaluation of ***various types of digital libraries***, for instance:

- digital libraries addressing different target audience such as children or undergraduate students;
- digital libraries addressing geographical locations, i.e. national digital libraries, rural digital libraries and state digital libraries;
- digital libraries addressing a particular subject area, such as computer science, medicine, mathematics, chemistry, etc.;
- digital libraries targeting a particular type of content, e.g. theses and dissertations, music digital libraries and video digital libraries.

2.7 Development of semantic digital libraries

Semantic Digital Libraries (SDLs) [5, 7, 8] are digital library systems that apply semantic technologies to achieve their specific goals. They make extensive use of metadata in order to support information retrieval and classification tasks. Ontologies play a major role to cope with the variety of problems caused by the structural differences of existing systems and the semantic differences of metadata standards. Within the context of SDLs, ontologies can be used to:

- organize bibliographic descriptions;
- represent and expose document contents;
- share knowledge amongst users.

There have been some successful efforts aiming to make use of ontologies and Semantic Web technology in digital libraries [7, 8]. For instance, JeromeDL[1] allows users to semantically annotate books, papers, and other resources. The Bricks[2] project aims to integrate existing digital resources into a shared digital memory. It relies on OWL DL in order to support, organize and manage metadata. Digital libraries within the biomedical domain store information related to methods, biomaterial, research statements, hypotheses, results, etc. Although the information is in the digital library, retrieving e.g. papers addressing the same topic and for which similar biomaterial has been used is not a trivial task.

Ontologies have shown to be useful for supporting the semantic annotation of scientific papers – and thereby facilitating information retrieval tasks. However, as ontologies are often incomplete, users should be able to provide additional metadata. Collaborative social tagging and annotation systems have recently gained attention within the research community – partly because of their

[1] http://www.jeromedl.org

[2] http://www.brickscommunity.org/

rapid and spontaneous growth and partly because of the need for structuring and classifying information. Collaborative social tagging is a considered activity in Web 2.0 because such sites use the Internet to harvest and utilize the collective intelligence.

3 Cloud Computing and Some Opportunities for Enhancement of Digital Library Services

Cloud computing is a technology that uses the Internet and central remote servers to maintain data and applications. Cloud computing allows consumers and businesses to use applications without installation and to access their personal files at any computer with Internet connection. This technology enables much more efficient computing by centralizing storage, memory, processing and bandwidth. There are many synonyms for Cloud computing: "On-demand computing", "Software as a service", "Information utilities", "The Internet as a platform", and others [9].

There are many reasons why this form of computing is attractive to organizations. It shifts the bulk of the responsibility for infrastructure support out to another vendor, and basically outsources all data center and software support to a company that specializes in Web-based computing. Using Cloud computing, one can share the server in many application procedures, realize the resource sharing and thus also reduce servers' quantity and achieve the effect of reducing the cost. Therefore, the utilization of Cloud computing in academic digital libraries will inevitably bring our work, the study and life to a greater efficiency.

With the rapid development of various information technologies, users' information requirements are increasingly personalized. More and more libraries advocate now user-centered services [10, 11]. So librarians should mine and study users' information requirements frequently. By establishing a pubic cloud among many academic libraries, it not only can conserve library resources but also can improve the users' satisfaction.

3.1 Development of semantic digital libraries

Library users still cannot access to the shared resources through an uniform access platform. However, with the adoption of Cloud computing in digital libraries, the integrated library resources support distributed uniform access interface. At the same time, the uniform access platform can promote library resources, guide and answer users' questions by using high-quality navigation. As a result, users can grip more information retrieval methods and make better use of library resources [12].

3.2 Integrated consulting services model

Today almost every university library can provide its users with network reference by Bulletin Board System (BBS) or Email. But with the constant improvement of users' demanding, integrated digital reference service came into being. And driven by Cloud computing, CDRS (Cooperative Digital Reference Service) can realize the sharing of technology, resources, experts and services of university libraries [13].

3.3 Real-time access services model

In the era of digital libraries, library users pay more attention to electronic journals, electronic databases and so on. This is really a big challenge for university libraries. But by introducing Cloud computing, university libraries can establish a shared public cloud jointly [14 – 16]. The shared cloud can have infinite storage capacity and computing power theoretically. It can bring obvious benefits to libraries.

On the one hand, allied libraries no longer consider the hardware cost. On the other hand, it will be possible to reduce the purchase of electronic database resources repeatedly among allied libraries. Meanwhile, users can visit the shared resources by any terminal equipment, such as PC, mobile phone etc., only if one can access to the Internet.

3.4 Knowledge service model

In the context of the knowledge economy, knowledge resource has become the main resource affecting productivity development. And university libraries are the main departments of storing, processing and spreading knowledge. So how to provide users with efficient transmission of information and knowledge services became urgent task for librarians today. The establishment of shared public cloud can save manpower and material resources greatly among university libraries [17]. Therefore, with the aid of Cloud computing, librarians won't have to maintain their own equipments or deal with consultations personally. And librarians will have more time and energy to offer the users their needed knowledge-based services but not only information sources.

3.5 Additional opportunities

Cloud computing is very closely related to the paradigm of Service Oriented Computing (SOC) [18]. Therefore, it is worth to command proper methods and tools for formal specification of cloud-supported services that assist service discovery, verification and optimization.

An approach to formal specification of Web services as an important component of SOC has been suggested in [19]. This approach can be easily

adapted to the Cloud computing paradigm. Together with formal specification of Cloud services, appropriate "direct" techniques for Cloud service discovery [20] are applicable in the sphere of enhancement of digital library services. It is also possible to adapt for the purpose some existing grid frameworks for library and E-learning services [21].

In addition, the possibility of integrating some popular open source software tools for creating digital repositories as e.g. DSpace[3] with platforms like DuraCloud [22] should also be considered. DuraCloud is an open source software platform which aims to maintain a fully integrated environment where services and data can be managed across multiple cloud providers and in this way a better support for data preservation, data transformation, and data access can be provided. So there are relatively simple ways to migrate from "traditional" implementations of digital libraries to cloud-based ones.

4 Conclusion

Library is not only a big knowledge source, its ultimate aim is to provide satisfactory services for everyone. So in the new era, library should improve itself constantly by adopting new information technologies. In particular, with the introduction of Cloud computing to university library, services of libraries will become more user-centric, more professional and more effective. The Cloud computing techniques and methods applied to digital libraries, can improve the utilization rate of resources to address the imbalance in development between regions and also can make more extensive use of Cloud computing to everyday work practice.

Acknowledgments. This work has been supported by the Bulgarian National Science Fund within Project ДДВУ 02/22/2010.

References

1. Shiri, A.: Digital Library Research: Current Developments and Trends. Library Review, Vol. 52, Issue 5 (2003), pp.198 – 202.
2. Pasquinelli, A. (Ed.): Digital Library Technology Trends. Sun Microsystems, 2002.
3. Seaman, D.: Aggregation, Integration, and Openness: Current Trends in Digital Libraries. International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society DLKC'04 (University of Tsukuba, Japan, March 2-5, 2004).
4. Semeraro, G. et al.: Intelligent Information Retrieval in a Digital Library Service. Proceedings of the First DELOS Network of Excellence Workshop on "Information Seeking, Searching and Querying in Digital Libraries" (Zurich, Switzerland, December 11-12, 2000), pp. 135 – 140.

[3] http://www.dspace.org

5. Castro, L., Giraldo, O., Castro, A.: Using the Annotation Ontology in Semantic Digital Libraries. In: A. Polleres, H. Chen (Eds.), Proceedings of the ISWC 2010 Posters & Demonstrations Track, Shanghai, China, 2010.
6. Nisheva-Pavlova, M.: Providing and Maintaining Interoperability in Digital Library Systems. Proceedings of the Fourth International Conference on Information Systems and Grid Technologies (Sofia, May 28-29, 2010), ISBN 978-954-07-3168-1, St. Kliment Ohridski Uiversity Press, 2010, pp. 200-208.
7. Kruk, S. et al.: The Role of Ontologies in Semantic Digital Libraries. 10th European Conference on Research and Advanced Technology for Digital Libraries, Alicante, Spain, 2006.
8. Kruk, S. et al.: JeromeDL – a Semantic Digital Library. In: J. Golbeck, P. Mika (Eds.), Proceedings of the Semantic Web Challenge Workshop at ISWC2007, Busan, South Korea, 2007.
9. Hayes, B.: Cloud Computing. Communications of the ACM, Vol. 51, No. 7 (2008), pp. 9-11.
10. Dillon, T., Wu, C., Chang, E.: Cloud Computing: Issues and Challenges. Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications, IEEE Computer Society, 2010, pp. 27-33.
11. Ullman, D., Haggerty, B.: Embracing the Cloud: Six Ways to Look at the Shift to Cloud Computing. EDUCAUSE Quarterly, Vol. 33, No. 2 (2010), http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/EmbracingtheCloudSixWaystoLook/206528 (accessed on April 15, 2012).
12. Liu, Y., Coleman, A.: Communicating Digital Library Services to Scientific Communities. LIBRES: Library and Information Science Research, Vol. 14 (1), March 2004.
13. Chen, Y. et al.: CALIS-based Cloud Library Services Platform Model. Advances in Information Sciences and Service Sciences, Vol. 3, No. 6, (July 2011), pp. 204-212.
14. Sanchati, R., Kulkarni, G.: Cloud Computing in Digital and University Libraries. Global Journal of Computer Science and Technology, Vol. 11, No 12 (2011), pp. 36-41.
15. Teregowda, P. et al.: Cloud Computing: A Digital Libraries Perspective. IEEE Cloud 2010, pp. 115-122.
16. Fox, R.: Library in the Clouds. OCLC Systems & Services, Vol. 25, Issue 3 (2009), pp.156 – 161.
17. Katz, R. (Ed.): The Tower and The Cloud. EDUCAUSE, 2008, ISBN 978-0-9672853-9-9, 295 pages.
18. Dimitrov, V.: Service Oriented Architecture. Sofia, Technologica, 2009, 208 pages (in Bulgarian).
19. Todorova, M.: Formal Specification of Web Services by Means of Generalized Nets with Stop-Conditions. Proceedings of the Fifth International Conference on Information Systems and Grid Technologies (Sofia, May 27-28, 2011), St. Kliment Ohridski Uiversity Press, 2011, ISSN 1314-4855, pp. 207-219.
20. Pashov, G., Kaloyanova, K.: Requirements to Cloud Service Discovery Systems. To appear in: Proceedings of the Sixth International Conference on Information Systems and Grid Technologies (Sofia – Gyolechica, Bulgaria, June 1-3, 2012).
21. Todorova, M.: Grid Framework for E-learning Services. Proceedings of the Fourth International Conference on Information Systems and Grid Technologies (Sofia, May 28-29, 2010), St. Kliment Ohridski Uiversity Press, 2010, ISBN 978-954-07-3168-1, pp. 153-163.
22. Kimpton, M., Payette, S.: Using Cloud Infrastructure as Part of a Digital Preservation Strategy with DuraCloud. EDUCAUSE Quarterly, Vol. 33, No. 2 (2010), http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/UsingCloudInfrastructureasPart/206548 (accessed on April 15, 2012).

Performance Evaluation of the Schemes for Dynamic Branch Prediction with ABPS Simulator

Milco Prisaganec[1], Pece Mitrevski[2], Nikola Rendevski[1]

[1]Faculty of administration and information systems management, University "St Kliment Ohridski" – Bitola, Macedonia
[2]Technical sciences, University "St Kliment Ohridski" – Bitola, Macedonia

Abstract. Modern processors in order to avoid the decline of performance due to stall the execution of the branch instruction in its architecture implement algorithms that perform branch prediction. They are intended to reduce the penalties that the microprocessor would be paid for time lost in resolving the branching instructions. Analyses show that the theoretically possible given predictor have an efficiency of 98% of correct predictions.
In this paper using a processor simulator with a certain type of branch predictor are measured performance offered by modern algorithms for prediction. Used the ***A**dvanced **B**ranch **P**rediction **S**imulator* (ABPS). ABPS simulator combines several types of two level adaptive branch predictor (GAg, GShare, PAg, PAp). Simulator allows you to change the basic parameters of the branch predictors that determine their accuracy. The tests used 17 professional testing programs. Shown are the dependence of the performance from hardware complexity of the branch predictor and the dependence on the type of coding on running application.
Research has shown that increasing the basic features of the branch prediction only to a certain point is justified. After that a major change to the complexity of the hardware gets a small percentage of increase accuracy of prediction. Also the results showed that the accuracy of prediction greatly depends on the type of application.

Keywords: branch prediction, processor, branch instruction

1. Introduction

A trend in the development of the new superscalar microprocessors is to achieve possible greater pipeline of instructions, vertically and horizontally, thereby to keep the economic cost in some limits. As a consequence of the parallel execution of instructions, a decline of microprocessor's performance occurs when it runs into the branch instruction. The analysis showed that the decline in performance due to the branch instructions may be between 30-40 percents. The basic strategy for reducing the performances while occurrence of the branch instructions is implementation of branch predictor in the superscalar microprocessor. Best results are obtained with the two-level adaptive predictors. In this paper by using an application will be simulated performance of 17 benchmarks through microprocessor which in its architecture contains one of the following branch predictors: GAg, GShare, PAg, PAp. During this the testing is performed by changing the hardware complexity of the predictor. The aim of the paper is to

show the impact of the complexity of the hardware from the branch predictors on the accuracy of predicted branches. It will show us what is the strategy for the design of the superscalar microprocessor's architecture and the implementation of the branch predictors.

A trend in the development of the new superscalar microprocessors is to achieve possible greater pipeline of instructions, vertically and horizontally, thereby to keep the economic cost in some limits. As a consequence of the parallel execution of instructions, a decline of microprocessor's performance occurs when it runs into the branch instruction. The analysis showed that the decline in performance due to the branch instructions may be between 30-40 percents. The basic strategy for reducing the performances while occurrence of the branch instructions is implementation of branch predictor in the superscalar microprocessor. Best results are obtained with the two-level adaptive predictors. In this paper by using an application will be simulated performance of 17 benchmarks through microprocessor which in its architecture contains one of the following branch predictors: GAg, GShare, PAg, PAp. During this the testing is performed by changing the hardware complexity of the predictor. The aim of the paper is to show the impact of the complexity of the hardware from the branch predictors on the accuracy of predicted branches. It will show us what is the strategy for the design of the superscalar microprocessor's architecture and the implementation of the branch predictors.

2. Dynamic branch prediction

The basic schemes for dynamic branch prediction that are based on the Smith's algorithm have more restrictions. The branch prediction is made on a limited history just for that branch instruction. The algorithm does not take into consideration the influence of the other branches in terms of what is currently predicted.

Experimentally is shown that there is a correlation between certain branches, and to increase the accuracy of prediction we use algorithms which store the history of the issue of the branch depending on the history of the correlative branches.

2.1. Two level adaptive schemes for branch prediction

In 1982, Yeh and Patt proposed two level adaptive techniques for branch prediction that uses flexible algorithm which is adapted to the changes of the context. In the Figure 2.1 there is a scheme of two-level adaptive branch predictor.

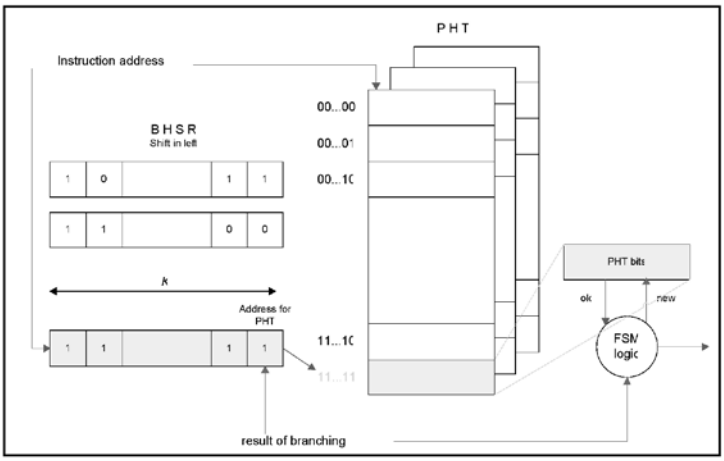


Figure 2.1 A two-level scheme for adaptive branch prediction

In the first level the scheme uses BHSR (**B**ranch **H**istory **S**hift **R**egister) which is indexed with *n* bits of the address of the branch instruction. The register contains the outcome of recent k executed branches. The second level contains a set of history tables of the branch that are identified as PHT (**P**attern **H**istory **T**able). The content of each entry of the Pattern History Table is the outcome of the recent branch instruction. The indexing of tables is made with n bits of the instruction branch address and relevant input from the table is indexed by the content of Branch History Shift Register.

The two level adaptive techniques for branch prediction is a framework where more variations can be implemented. There are three options for implementation of Branch History Shift Register, and the outcome of recent *k* executed branches is being saved, which means:

- If the recent *k* branches are saved (G - Global);
- If the recent *k* appearances of the same branch instruction are saved (P - Per address);
- If the recent *k* appearances of branch instructions from the same set are saved (S - per Set).

Also for PHT structure there are three options for implementation, as follows:

- g - global, that uses a PHT table for all branch predictions;
- p - local, when each PHT table is used for unique branch, and
- s - shared, when there exists only one PHT table for each set of branch instructions.

All possible implementations on two level adaptive branch prediction depend on

the way the history is saved in the first level (G, P, S) and the way the tables of the second level are connected with the history information obtained from the first level (g, p, s). The possible variations of two level adaptive branch predictors are shown in the Figure 2.2.

		Scraps of the second level history		
		g	p	s
Scraps of the first level history	G	GAg	GAp	GAs
	P	PAg	PAp	PAs
	S	SAg	SAp	SAs

Figure 2.2 Implementations of two level adaptive branch predictions

Using a two-level adaptive algorithm for branch prediction achieves high prediction accuracy of 95%.

2.2. Software package for simulation

For benchmarking and performance evaluation of schemes for dynamic branch prediction in this paper is used the simulator Advanced Branch Prediction Simulator (ABPS) [9], and its authors are Calborean Horia, Crapciu Adrian and Radu Ciprian. The main goal of the simulator is to allow easy introduction with the problems of the branch prediction.

ABPS simulator contains several types of two-level adaptive branch predictors (GAg, GShare, PAg, PAp). For each branch predictor there is a possibility for changing the parameters that affect the accuracy of prediction. Moreover, there is a possibility to change the size of the global history (GH-Global History), the size of the PHT (Pattern History Table), to perform a change in the number of bits of program counter (PC-Program Counter), which will be used when indexing and the number of bits used for saving the outcome of recent branch.

The application is developed in Java programming language, which allows its use on many different platforms of operating systems.

ABPS allows using two types of benchmarks, as follows:

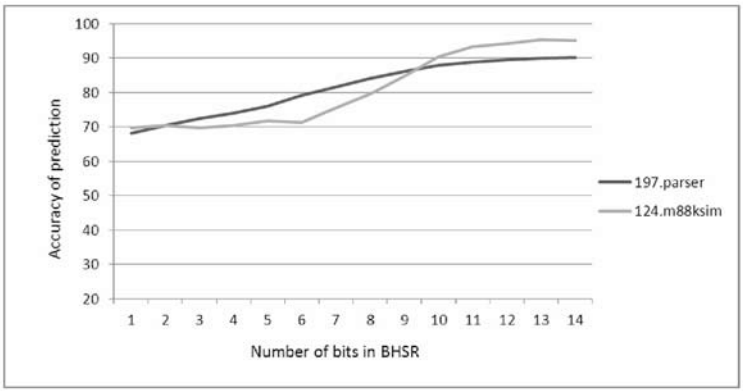
- Stanford benchmarks - is an older version, but is dedicated to educate students and easily overcome of the problems of branch predictions;
- SPEC 2000 benchmarks - is dedicated for professional use, and contains 17 different benchmarks – with one million branch instructions in each. Benchmarks are made by many authors and as an algorithm are used in applications from different areas.

2.3 Overview of results and analysis

During the Benchmarking are used all 17 benchmarks through algorithms for dynamic branch prediction in the application. During each simulation parameters of branch prediction are changed from which depends the accuracy of prediction.

2.3.1 Simulation of GAg branch predictor

GAg branch predictor in the first level has one global registry in which is the recent *k* outcomes of the branch instructions. Also in the second level there is only one PHT table which stores the last outcome of all branch instructions. The indexing on the entrances of PHT table is done with the contents of the global register. During the benchmarking the changes are done on the number of bits of BHSR thereby increasing the number of inputs in GPHT. The benchmarking of two benchmarks is shown in graph 2.1.



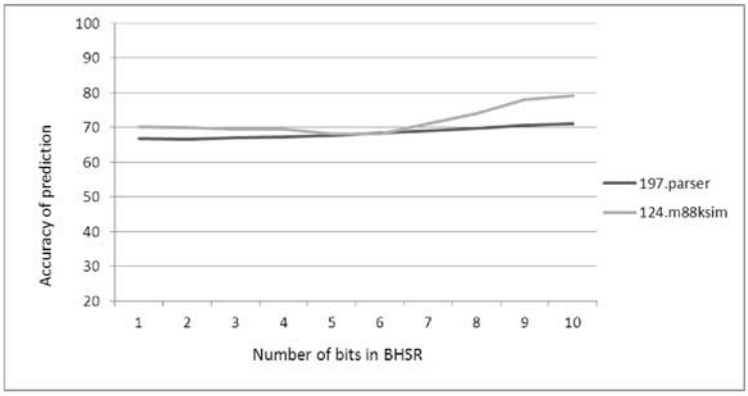
Graph 2.1: Accuracy of prediction in GAg

The result shows that by increasing the number of bits in the global registry the percentage of accurate branch predictions is increased. But on the graph can be seen that after a certain number of bits (12 bits displayed benchmarks) the growth in the number of accurate branch predictions is negligible. Irregularities in the growth of curves of the particular benchmark 124.m88ksim arise from the existence of a global register and a PHT table for all instructions, so very easy can cause interference or indexing a wrong entry in the PHT.

2.3.2 Simulation on GShare branch predictor

Structurally the scheme GShare branch predictor is the same as in the previous predictor. And in it there is a single global register in which are stored the recent

k outcomes of the branch instructions and one GPHT table. The only difference is in the indexing of the inputs of GPHT i.e. the address is obtained by XOR operation between BHSR and the address of the branch instruction. During the benchmarking it is set to be taken the two low bits of the address of instruction and changes are made to the number of bits in the global registry and therefore the number of inputs in GPHT. The results of the benchmark for two benchmarks are shown in graph 2.2.

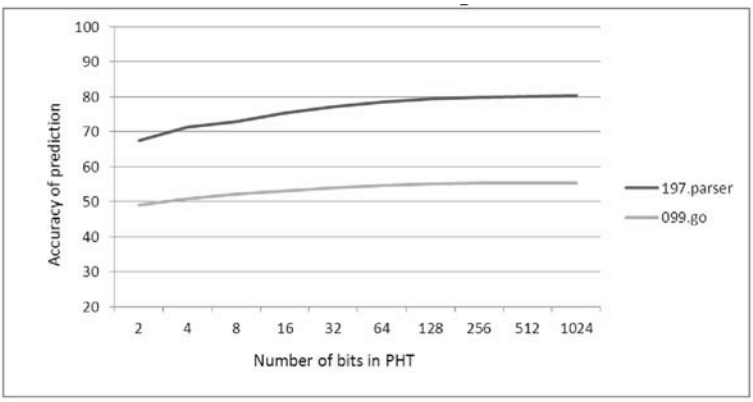


Graph 2.2: Accuracy of prediction in GShare

From the graph can be reached the same conclusions as before that by increasing the number of bits in the global registry and thus increasing the number of entries in the table GPHT the percentage of accurate branch predictions grow. But also here we notice that it occurs only to certain number of bits in the global registry and after that the growth of accurate predictions is negligible. The irregularities that appear in the curve of the benchmark 124.m88ksim again are due to the interference that occurs due to indexing of one global PHT table for all branch instructions.

2.3.3 Simulation of PAg branch predictor

During the benchmarking at PAg the BHT predictor has 8 inputs that are indexed with three bits of the address of instruction. The number of entries in the PHT table increases to the size of 1024 entries. On the graph 2.3 are shown the results of the two benchmarks i.e. the number of accurate branch predictions in percentage depending on the size of the PHT table which is shared for this predictor.

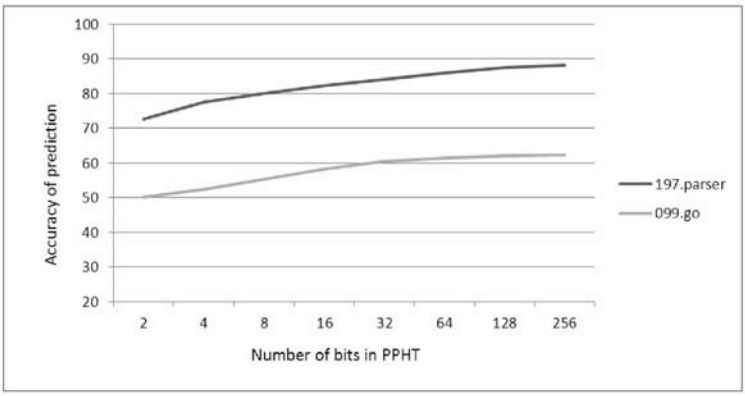


Graph 2.3: Accuracy of prediction in Pag

The results show that with increasing the entries in PHT table, the number of accurate branch predictions increases. But it is obvious that by increasing the number of entries in the PHT table more than 256 the growth of the accurate branch predictions is insignificant.

2.3.4 Simulation of PAp branch predictor

In this scheme for prediction it is characteristic that for each branch instruction there is a special BHSR register and a special PHT table. During the benchmarking the size of PBHT table is 256 entries and doesn't change, and as a variable is the size of the PPHT table. Under those conditions are used all 17 benchmarks through the branch predictor. The results for two benchmarks are shown on the graph 2.4.



Graph 2.4: Accuracy of prediction in Pap

In addition, in both cases can be detected the same conclusion as in the previous predictors that with increasing the input of PHT table the precision on the exact branch predictions grows but to a certain limit after which the increase is negligible. And in both algorithms where we have separately BHSR for each branch instruction i.e. and a special PHT table the accuracy of branch predictions is quite high. The results obtained by counting the accurate predictions when certain benchmarks are executed simulating that the microprocessor has two-level adaptive branch predictor and with their analysis have come to the following conclusions:

- Increasing the parameters in predictors leads to an increase in the percentage of accurate predict outcome branch instructions. But also in all of them was proved that there is an upper limit after which increased accuracy in prediction is negligible or doesn't exist. Accuracy in percentage ranges from 50% - 98% depending on the predictor and his complexity.
- For all branch predictors their accuracy is different for different benchmarks. In other words the number of correctly outcome predictions at the branch instructions depends on the nature of the encoding application that the microprocessor executes. For some algorithms for prediction the percentage of accurate branch predictions is in the interval 65% - 98%.
- Branch predictions have different accuracy when it comes to the same benchmark and when their parameters in the first and second level are close to each other.

When designing the modern superscalar microprocessor above listed conclusions indicate the need for implementation of the branch predictors in the architecture. In this the analysis showed that it is possible to achieve accuracy of 98% when making speculative decision during using complex hardware. But it is not viable commercially, so it is necessary to find the optimal solution. In other words, using multiple types of branch predictions whose complexity would be justified commercially, and on behalf of that consciously reducing the performances on the superscalar microprocessor.

3. Conclusion

The analysis showed that it is possible to get very high percentage of accurate branch predictions but it is necessary to use expensive hardware solution which is certainly not the purpose of the future development of the microprocessors. It is proved that the performance of the branch predictors depends also on the type of coding the program that is executing through the microprocessor and the use of only one algorithm for prediction would mean dependence on the performances of the microprocessor of the application executed. So the goal in the design of the modern superscalar microprocessors is to achieve an optimal solution cost /

performance. Because of that it is better to use two or more algorithms to predict who will offer an optimal solution, which cheap implementation consciously will reduce the performances of the microprocessor. The microprocessor would group several different algorithms for branch prediction in order to cover different encodings of those applications that microprocessor will perform but mutually to cover the weaknesses of certain branch predictors.

REFERENCES

[1] John Paul Shen –Mikko H. Lipasti: *Modern processor design-fundamentals of superscalar processors*, McGraw – Hill publishing company design, 2005

[2] Mitrevski, P., Gusev, M., "*On the Performance Potential of Speculative Execution Based on Branch and Value Prediction*", *International Scientific Journal Facta Universitatis, Series: Electronics and Energetics*, Vol. 16, No. 1, ISSN: 0353-3670, pp. 83-91, 2003

[3] Milco Prisaganec: *Performance evaluation of schemes dynamic branch prediction*, Master's Thesis, 2011

[4] Tse-Yu Yeh – Yale N. Patt: *Alternative implementations of two-level adaptive branch prediction*, The university of Michigan, 1992

[5] Gusev, M., Mitrevski, P., "Modeling and Performance Evaluation of Branch and Value Prediction in ILP Processors", *International Journal of Computer Mathematics*, Vol. 80, No. 1, ISSN: 0020-7160 (print); ISSN: 1029-0265 (online), pp. 19-46, 2003

[6] Yeh, T.-Y.; Patt, Y. N.: *"Two-Level Adaptive Training Branch Prediction"* ,Proceedings of the 24th annual international symposium on Micro architecture, Albuquerque, New Mexico, Puerto Rico: ACM, 1991

[7] Tse-Yu Yeh – Yale N. Patt : *"A Comparison of Dynamic Branch Predictors that use Two Levels of Branch History",* The 20th Annual International Symposium on Computer Architecture pp. 257 - 266, May 16 - 19, 1993, San Diego, California.

[8] Steven G. B.-Egan C. Shim-W. Vintan L.: *A cost-effective two-level adaptive branch predictor*, University of Hertfordshire, Seoul National University of Technology, Romania

[9] http://abps.sourceforge.net/

Data analysis for next generation sequencing - parallel computing approaches in *de novo* assembly algorithms

Dimitar Vassilev[1*], Valeria Simeonova[2], Milko Krachunov[2], Elena Todorovska[1], Maria Nisheva[2], Ognyan Kulev[2], Deyan Peychev[1], Peter Petrov[2], Dimitar Shiyachki[2], Irena Avdjieva[1]

[1] *Bioinformatics group, AgroBioInstitute, Sofia 1164, 8 Dragan Tsankov Blvd*

[2] *Faculty of Mathematics and Informatics, Sofia University "St.Kliment Ohridski"*

Corresponding author: jim6329@gmail.com

Abstract

The new parallel sequencing technologies produce gigabases of genome information in just a few days bring with them new problems for data storage and processing. Sequencing technologies have applications in human, plant and animal genome studies, metagenomics, epigenetics, discovery of non-coding RNAs and protein binding sites. There are two major problems in next generation sequencing (NGS) data processing: algorithms for alignment of sequences (for which exists a reference sequence) and algorithms for *de novo* genome (sequence) assembly (for which no reference sequence is available). Different factors define the choice of better algorithmic solution: cost, reads length, data volume, rate of data generation). As a result the particular bioinformatics solution depends on the biological application and on the type of sequencing technology used to generate the data. All the technologies have their strengths and weaknesses and limits of their performance for providing error free sequenced data.

The goal of the paper is to discuss the problems concerning methods, algorithms and software solutions for data analysis as error mapping, assembly and annotation in *de novo* sequencing studies and to estimate the amount of change we can introduce to the raw output data without losing too much information. The opportunities for parallelization of some of the algorithms and procedures used in sequence alignment and assembly are also in the focus of the paper.

Keywords: next generation sequencing, *de novo* genome assembly, parallel computing

1. Introduction

Next-generation sequencing (NGS) technologies are demonstrating breath-taking revolutionary changes in research in life sciences [1]. These high-

throughput technologies are widely applied in many research and domains and practices including metagenomics [2], detection of SNPs [3] and genomic structural variants [4,5] in a population, DNA methylation studies [6], analysis of mRNA expression [7], cancer genomics [8] and personalized medicine [9]. Some applications (e.g. metagenomics, plant and animal genomics) require *de novo* sequencing of a sample [10], while many others (e.g. variant detection, cancer genomics) require re-sequencing. For all of these applications, the vast amount of data produced by sequencing runs poses many computational challenges [11].

In re-sequencing, a reference genome is already available for the species (e.g. the human genome) and one is interested in comparing short reads obtained from the genome of one or more donors (individual members of the species) to the reference genome. Therefore, the first step in any kind of analysis is the mapping of short reads to a reference genome. This task is complicated by many factors, including genetic variation in the population, sequencing error, short read length and the huge volume of short reads to be mapped.

The vast number of short reads produced by these techniques, however, poses significant computational challenges. The first step in many types of genomic analysis is the mapping or aligning of short reads to a reference genome, or development special algorithmic strategies if the reference genomes are not available. There is a number of research and business groups who have developed dedicated algorithms and software packages to perform this function. As the developers of these packages optimize their algorithms with respect to various considerations, the relative merits of different software packages remain unclear. However, for scientists who generate and use NGS data for their specific research projects, an important consideration is how to choose or to elaborate a software solution that is most suitable for their application [12].

Along with that the advent of short-read sequencing machines gave rise to a new generation of IT application molecular biology studies. The algorithms and software developed for analyzing of short-read data is focused mainly on the error/mutation detection and assembly of the already "clean" data. The constant emerging of new and improved algorithms for *de novo* whole-genome assembly and annotation from next-generation sequencing data explicitly emphasizes the unambiguous impact of algorithmic techniques and their software solutions for practical implementation.

So far, many algorithms have been developed to overcome these challenges and these algorithms have been made available to the scientific community as software packages [13]. Currently available software packages for short read alignment include Bowtie [14], SOAP [15,16], BWA [17], mrFAST [4], mrsFAST [18], Novoalign [19] and SHRiMP [20].

Nevertheless, the sequences are coming, and the bioinformatics community needs to act quickly to keep pace with the expected flood of raw sequence data. In

this review, we describe the challenges facing those who use genome assembly and annotation software and review the initial efforts to develop new bioinformatics software for *de novo* assembly of short read sequencing (SRS) technology.

2. Next Generation Sequencing Technologies and Data Production

The development of the new massively parallel sequencing technologies has sprung from recent advances in the field of nanotechnology, from the availability of optical instruments capable of reliably detecting and differentiating millions of sources of light or fluorescence on the surface of a small glass slide and from the ingenious application of classic molecular biology principles to the sequencing problem. Another important consideration is that, in the context of an already available genome sequence, many problems such as the identification of single nucleotide polymorphisms (SNPs) does not require the generation of ever longer sequence reads, because most possible words of length >25 or 30 only occur at most once even in relatively large genomes allowing, for the most part, unambiguous assignment of even the shortest reads to a locus of origin in a reference genome. Thus, available NGS technologies produce large numbers of short sequence reads and are used both in *de novo* sequencing and in re-sequencing applications.

Currently available next-generation sequencers rely on a variety of different chemistries to generate data and produce reads of differing lengths, but all are massively parallel in nature and present new challenges in terms of bioinformatics support required to maximize their experimental potential.

Three distinct NGS platforms have already attained wide distribution and availability. Some characteristics of their throughput, read-lengths and costs (at the time of writing) are presented in Table1.

Table 1. Performances and features of the major next-generation sequencing platforms (single-end reads**)**

Technology	Roche 454			Illumina		ABI Solid		
Platform	*GS 20*	*FLX*	*Ti*	*GA*	*GA II*	*1*	*2*	*3*
Reads (M)	0,5	0,5	1	28	100	40	115	400
Read length	100	200	350	35	75	25	35	50
Run time (d)	0,2	5	0,3	0,4	4,5	6	5	6-7
Images (TB)	0,01	0,01	0,03	0,5	1,7	1,8	2,5	3

A common thread for each of these technologies over the last years has been continuous improvement in performance (increased numbers and lengths of reads and consequent reduction in costs per base sequenced), it is therefore anticipated that the figures provided will rapidly become outdated, however, they serve to

illustrate that the Roche454 technology [21] already provides a realistic substitute for many applications of conventional Sanger sequencing at greatly reduced cost, while the Illumina Genome Analyser [22] and ABISOLiD [23] platforms generate an order of magnitude more reads of (relatively) reduced length, characteristics that, as we will see, render them, for now, more suitable for other applications. The afore-mentioned methods all rely on a template amplification phase prior to sequencing. However, the available Helicos technology [24] avoids the amplification step and provides sequence data for individual template molecules, minimizing the risk of introducing substitutions during amplification. In principle bioinformatics algorithmic approaches developed for the analysis of data generated by the Illumina GA and ABISOLiD platforms should also be suitable for data generated by the Helicos method, as all three platforms provide reads of comparable lengths, however there are many different details for every particular data analysis. Finally, other methods, based on either nanopore technology or tunneling electron microscopy have been proposed (for reviews see [25, 26]). Detailed information on the performance of such approaches is not yet wide available, although it is hoped that they could yield individual reads of lengths measured in megabases. Given that such methods remain broadly inaccessible at the present time, and that the nature of data generated should be fundamentally different from those provided by available platforms. Potential bioinformatics developments connected to these methods are considered to be major challenge for future generations sequencing technologies.

3. Sequence properties and algorithmic challenges

NGS technologies typically generate shorter sequences with higher error rates from relatively short insert libraries [28,29,30]. For example, one of the most commonly used technologies, Illumina's sequencing by synthesis, routinely produces read lengths of 75 base pairs (bp) from libraries with insert sizes of 200–500 bp. It is, therefore, expected that assembly of longer repeats and duplications will suffer from this short read length. Similar to the whole-genome shotgun sequence (WGS) assembly algorithms that use capillary-based data such as the Celera assembler [27], the predominant assembly methods for short reads are based on de Bruijn graph and Eulerian path approaches [31, 32], which however have some problems in assembling complex regions of the genome. As argued by groups that presented various implementations of this approach (for example, the algorithms named EULER-USR [33], AbySS [34] and SOAPdenovo [35]), paired-end sequence libraries with long inserts help to ameliorate this bias. However, even the longest currently wide available inserts (<17 kilobases with Roche 454 sequencing [28]) are insufficient to bridge across regions that harbor the majority of recently duplicated human genes. Criticisms of WGS assembly

algorithms and characterization of various types of errors associated with them as well as requirements for better assemblies are widely discussed [28,29, 36].

4. Advances in algorithms and software tools for alignment of short read sequencing data

The rapid development of new sequencing technologies substantially extends the scale and resolution of many biological applications, including the scan of genome wide variation [37], identification of protein binding sites (ChIP-seq), quantitativeanalysis of transcriptome (RNA-seq) [38], the study of the genome-wide methylation pattern [39] and the assembly of new genomes or transcriptomes [40]. Most of these applications take alignment or *de novo* assembly as the first step; even in *de novo* assembly, sequence reads may still need to be aligned back to the assembly as most large-scale short-read assemblers [35,36] do not track the location of each individual read. Sequence alignment is therefore essential to nearly all the applications of new sequencing technologies.

Most of fast alignment algorithms construct auxiliary data structures, called indices, for the read sequences or the reference sequence, or sometimes both. Depending on the property of the index, alignment algorithms can be largely grouped into three categories: algorithms based on hashtables, algorithms based on suffix trees and algorithms based on merge sorting. The third category only consists of Slider[41] and its descendant SliderII [42]. This review will therefore focus on the first two categories.

Table 2. Some popular short read alignment software programs

Program	Algorithm	SOLiD	454 Roche [a]	Gapped	Illumina (PE[b])	Q[c]
Bfast(BLAST)	hashing.ref	Yes	No	Yes	Yes	No
Bowtie	FM index	Yes	No	No	Yes	Yes
BWA	FM index	Yes [d]	Yes [e]	Yes	Yes	No
MAQ	hashing reads	Yes	No	Yes [f]	Yes	Yes
Mosaik	hashing ref.	Yes	Yes	Yes	Yes	No
Novoalign [g]	hashing ref.	No	No	Yes	Yes	Yes

[a]Work well for Sanger and 454 reads, allowing gaps and clipping. [b]Paired end mapping. [c]Make use of base quality in alignment. [d]BWA trims the primer base and the first color for a color read. [e]Long-read alignment implemented in the BWA-SW module. [f]MAQ only does gapped alignment for Illumina paired-end reads. [g]Free executable for non-profit projects only.

4.1. Algorithms based on hash tables

The idea of hash table indexing can be tracked back to BLAST [43, 44].

All hash table based algorithms essentially follow the same seed-and-extend paradigm. BLAST keeps the position of each ***k-mer*** (k_{j11} by default) subsequence of the query in a hash table with the ***k-mer*** sequence being the key, and scans the database sequences for ***k-mer*** exact matches, called seeds, by looking up the hash table. BLAST extends and joins the seeds first without gaps and then refines them by a Smith–Waterman alignment [45,46]. It outputs statistically significant local alignments as the final results. The basic BLAST algorithm has been improved and adapted to alignments of different types. Nevertheless, the techniques discussed below focus on mapping a set of short query sequences against a long reference genome of the same species.

Improvement on seeding: spaced seed. BLAST seeds alignment with 11 consecutive matches by default. Ma et al. [47] discovered that seeding with non-consecutive matches improves sensitivity. For example, a template '111010010100110111' requiring 11 matches at the '1' positions is 55% more sensitive than BLAST's default template '11111111111' for two sequences of 70% similarity. A seed allowing internal mismatches is called spaced seed; the number of matches in the seed is its weight. Eland (A.J. Cox) was the first program that utilized spaced seed in short-read alignment. It uses six seed templates spanning the entire short read such that a two-mismatch hit is guaranteed to be identified by at least one of the templates, no matter where the two mismatches occur. SOAP [48] adopts almost the same strategy except that it indexes the genome rather than reads.

It is memory demanding to hold in RAM a hash table with ***q*** larger than 15. Many other aligners [49,50,51] also use spaced seed with different templates designed specifically for the reference genome and sensitivity tolerances, making spaced seed the most popular approach for short-read alignment.

A potential problem with consecutive seed and spaced seed is they disallow gaps within the seed. Gaps are usually found afterwards in the extension step by dynamic programming, or by attempting small gaps at each read positions [48, 52]. The ***q-gram*** filter, as is implemented in SHRiMP [53] and RazerS [54], provides a possible solution to building an index natively allowing gaps.

Improvements on seed extension. Due to the use of long spaced seeds, many aligners do not need to perform seed extension or only extend a seed match without gaps, which is much faster than applying a full dynamic programming. Nonetheless, several improvements over BLAST have been made regarding on seed extension. A major improvement comes from the recent advance in accelerating the standard Smith–Waterman with vectorization. The basic idea is to parallelize alignment with the CPU SIMD instructions such that multiple parts of a query sequence can be processed in one CPU cycle. Using the SSE2 CPU instructions implemented in most latest x86 CPUs, [55] derived a revised Smith–

Waterman algorithm that is over 10 times faster than the standard algorithm. Novoalign (http://novocraft.com), CLC Genomics Workbench (http://clcbio.com/index.php?id=1240) and SHRiMP are known to make use of vectorization.

Another improvement is achieved by constraining dynamic programming around seeds already found in the seeding step [56,57,58]. Thus unnecessary visits to cells far away from seed hits in iteration are greatly reduced. In addition, Myers [59] found that a query can be aligned in full length to an L-long target sequence with up to k mismatches and gaps in ***O(kL)*** time, independent of the length of the query. These techniques also help to accelerate the alignment when dynamic programming is the bottleneck.

4.2. Algorithms based on suffix/prefix tries

All algorithms in this category essentially reduce the inexact matching problem to the exact matching problem and implicitly involve two steps: identifying exact matches and building inexact alignments supported by exact matches. To find exact matches, these algorithms rely on a certain representation of suffix/prefix trie, such as suffix tree, enhanced suffix array [60] and FM-index [61]. The advantage of using a trie is that alignment to multiple identical copies of a substring in the reference is only needed to be done once because these identical copies collapse on a single path in the trie, whereas with a typical hash table index, an alignment must be performed for each copy.

5. The challenge of assembly algorithms and software solutions for next generation sequencing data

According to [28] - an assembly is a hierarchical data structure that maps the sequence data to a putative reconstruction of the target. It groups reads into contigs and contigs into scaffolds. Contigs provide a multiple sequence alignment of reads plus the consensus sequence. The scaffolds, sometimes called super contigs or metacontigs, define the contig order and orientation and the sizes of the gaps between contigs. Scaffold topology maybe a simple path or a network. Most assemblers output, in addition, a set of unassembled or partially assembled reads. The most widely accepted data file format for an assembly is FASTA, wherein contig consensus sequence can be represented by strings of the characters ***A, C, G, T,*** plus possibly other characters with special meaning. Dashes, for instance, can represent extra bases omitted from the consensus but present in a minority of the underlying reads. Scaffold consensus sequence may have ***N's*** in the gaps between contigs. The number of consecutive ***N's*** may indicate the gap length estimate based on spanning paired ends. Assemblies are measured by the size and accuracy of their contigs and scaffolds. Assembly size is usually given by statistics including maximum length, average length, combined total length, and

N50. The contig ***N50*** is the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly. The ***N50*** statistics for different assemblies are not comparable unless each is calculated using the same combined length value. Assembly accuracy is difficult to measure. Some inherent measure of accuracy is provided by the degrees of mate-constraint satisfaction and violation [62].

5.1. Basic Principles of Assembly

For the majority of traditional assembly programs the basic scheme is the same, namely the overlap-layout-consensus approach. Essentially it consists of the following steps [63,64]:

- Sequence and quality data are read and there ads are cleaned.
- Overlaps are detected between reads. False overlaps, duplicate reads, chimeric reads and reads with self-matches (including repetitive sequences) are also identified and left out for further treatment.
- The reads are grouped to form a contig layout of the finished sequence.
- A multiple sequence alignment of the reads is performed, and a consensus sequence is constructed for each contig layout (often along with a computed quality value for each base).
- Possible sites of mis-assembly are identified by combining manual inspection with quality value validation.

5.2. General Assembly Differences

When different assemblers try to piece the DNA puzzle together they essentially work from the same input, but the assemblers differ in the way they utilize the sequence information, and in the way this is combined with additional information. In general the differences fall in the following categories.

- **Overlaps:** A lot of different methods are used to find potential overlaps between sequences. Some are based on BLAST (e.g. Gene Distiller by Gilchrist et al., [65] and [29]); while other assemblers use various other methods to find similarities between reads.

- **Additional information:** Depending on how the sequence reads are produced some additional information might be available. This information might consist of read pair information, BAC clone information, base quality information, etc. Some assemblers use this data to impose additional structure on the assembly of the sequences (e.g. GigAssembler International Human Genome Sequencing Consortium, [66]).

- **Short read assembly:** *De novo* assembly of the short reads generated from next generation sequencing platforms is still challenging. While assemblers have been developed and applied to assemble bacterial genomes successfully [67, 68] on

larger genomes the assembly is performed in general by mapping the microreads to reference genomes. The major next generation sequencing platforms all have built-in software to handle this task, e.g. GS Reference mapper, Gerald for Solexa. In SOLiD systems the mapping tool "mapreads" converts reference sequences into color space and perform the mapping in color space. However the built-in software very often gives some unpredictable biases and the new bioinformatics developments are oriented towards pipelines and suits for a custom designed software solutions for genome and transcriptome assembly.

DNA sequencing technologies share the fundamental limitation that read lengths are much shorter than even the smallest genomes. WGS (Whole Genome Sequencing) overcomes this limitation by over-sampling the target genome with short reads from random positions. Assembly software reconstructs the target sequence and the assembly software is challenged by repeat sequences in the target. Genomic regions that share perfect repeats can be indistinguishable, especially if the repeats are longer than the reads. For repeats that are inexact, high-stringency alignment can separate the repeat copies. Careful repeat separation involves correlating reads by patterns in the different base calls they may have [69]. Error tolerance leads to false positive joins. This is a problem especially with reads from inexact (polymorphic) repeats. False-positive joins can induce chimeric assemblies. In practice, tolerance for sequencing error makes it difficult to resolve a wide range of genomic phenomena: polymorphic repeats, polymorphic differences between non-clonal asexual individuals, polymorphic differences between non-inbred sexual individuals, and polymorphic haplotypes from one non-inbred individual. If the sequencing platforms ever generate error-free reads at high coverage, assembly software might be able to operate at 100% stringency.

NGS assembly is complicated by the computational complexity of processing larger volumes of data. For efficiency, all assembly software relies to some extent on the notion of a ***k-mer***. This is a sequence of ***k*** base calls, where ***k*** is any positive integer. In most implementations, only consecutive bases are used. Intuitively, reads with high sequence similarity must share ***k-mers*** in their overlapping regions, and shared ***k-mers*** are generally easier to find than overlaps. Fast detection of shared ***k-mer*** content vastly reduces the computational cost of assembly, especially compared to all-against-all pairwise sequence alignment. A tradeoff of ***k-mer*** based algorithms is lower sensitivity, thus missing some true overlaps. The probability that a true overlap spans shared ***k-mers*** depends on the value of ***k***, the length of the overlap, and the rate of error in the reads. An appropriate value of ***k*** should be large enough that most false overlaps don't share ***k-mers*** by chance, and small enough that most true overlaps do share ***k-mers***. The choice should be robust to variation in read coverage and accuracy [28].

NGS assembly algorithms, and their implementations, are typically complex. Assembly operation can require high-performance computing platforms for

large genomes. Algorithmic success can depend on pragmatic engineering and heuristics, that is, empirically derived rules of thumb. Heuristics help overcome convoluted repeat patterns in real genomes, random and systematic error in real data, and the physical limitations of real computers.

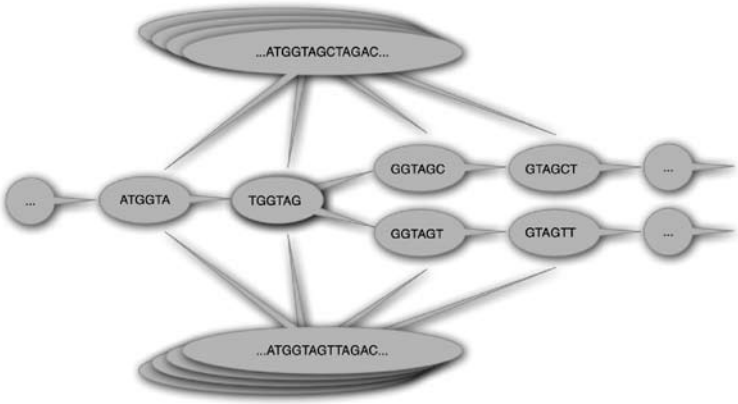


Fig 2. Graph example. An examples of how a graph is constructed. Two reads are mapped onto the different ***k-mer*** nodes (***k = 6*** in this example), and edges between the nodes are determined by the reads. The presence of a nucleotide difference (e.g. sequencing error, SNP, etc.) between the two reads cause the graph to split up, thus causing an ambiguity in the sequence.

A somewhat related issue is how the sequences are cleaned of errors and contaminant sequences (i.e. vector sequences, repeat sequences, etc.). While this can essentially be considered separately and independently from the assembly itself, some assemblers incorporate cleaning in the way they process the reads [28, 29]

5.1. Graph Algorithms for Assembly

NGS assemblers could be classified into three categories, all based on graphs. The Overlap/Layout/Consensus (OLC) [40, 70, 28] methods rely on an overlap graph. The de Bruijn Graph (DBG) methods use some form of ***k-mer*** graph [71, 32]. The greedy graph algorithms may use OLC or DBG [28]. A graph is an abstraction used widely in computer science. It is a set of nodes plus a set of edges between the nodes. Nodes and edges may also be called vertices and arcs, respectively. If the edges may only be traversed in one direction, the graph is known as a directed graph. The graph can be conceptualized as balls in space with arrows connecting them. Importantly, each directed edge represents a connection from one source node to one sink node. Collections of edges form paths that visit nodes in some order, such that the sink node of one edge forms the source node for any subsequent nodes. A special kind of path, called a simple path, is one that contains only distinct nodes (each node is visited at most once). A simple

path may not intersect itself, by definition, and one may additionally require that no other paths intersect it. The nodes and edges may be assigned a variety of attributes and semantics.

An overlap graph represents the sequencing reads and their overlaps [70, 28]. The overlaps must be pre-computed by a series of (computationally expensive) pair-wise sequence alignments. Conceptually, the graph has nodes to represent the reads and edges to represent overlaps. In practice, the graph might have distinct elements or attributes to distinguish the 5′ and 3′ ends of reads, the forward and reverse complement sequences of reads, the lengths of reads, the lengths of overlaps, and the type of overlap (suffix-to-prefix or containment). Paths through the graph are the potential contigs, and paths can be converted to sequence. Paths may have mirror images representing the reverse complement sequence. There are two ways to force paths to obey the semantics of double-stranded DNA. If the graph has separate nodes for read ends, then paths must exit the opposite end of the read they enter. If the graph has separate edges for the forward and reverse strands, then paths must exit a node on the same strand they enter.

The de Bruijn graph was developed outside the realm of DNA sequencing to represent strings from a finite alphabet [73]. The nodes represent all possible fixed-length strings. The edges represent suffix-to-prefix perfect overlaps. De Brujn graphs were first brought to bioinformatics as a method to assemble ***k-mers*** generated by sequencing by hybridization [72, 74]; this method is very similar to the key algorithmic step of today's short-read assemblers.

A ***k-mer*** graph is a form of de Bruijn graph. Its nodes represent all the fixed-length subsequences drawn from a larger sequence. Its edges represent all the fixed-length overlaps between subsequences that were consecutive in the larger sequence [32]. In one formulation [72], there is one edge for the ***k-mer*** that starts at each base (excluding the last ***k-1*** bases). The nodes represent overlaps of ***k-1*** bases. Alternately [32], there is one node representing the ***k-mer*** that starts at each base. The edges represent overlaps of ***k-1*** bases. By construction, the graph contains a path corresponding to the original sequence The path converges on itself at graph elements representing ***k-mers*** in the sequence whose multiplicity is greater than one.

A repeat graph is an application of the ***k-mer*** graph [75]. It provides a succinct graph representation of the repetitiveness of a genome. Nodes and edges are drawn from an assembled reference sequence. Whereas non-repetitive genomic sequence would induce a single path through the graph, repeats induce convergence and divergence of paths, as well as cycles. Repeat graphs can be used to identify and catalog repeats [76].

In general, branching and convergence increases graph complexity, leading to tangles that are difficult to resolve. Much of the complexity is due to repeats in the target and sequencing error in the reads.

In the graph context, assembly is a graph reduction problem. Most optimal graph reductions belong to a class of problems, called ***NP-hard***, for which no efficient solution is known [77]. Therefore, as it was mentioned above, assemblers rely on heuristic algorithms and approximation algorithms to remove redundancy, repair errors, reduce complexity, enlarge simple paths and otherwise simplify the graph.

5.2.Types of assemblers

Greedy Graph-based Assemblers. The first NGS assembly packages used greedy algorithms. These have been reviewed well elsewhere [78, 79]. The greedy algorithms apply one basic operation: given any read or contig, add one more read or contig. The basic operation is repeated until no more operations are possible. Each operation uses the next highest-scoring overlap to make the next join. The scoring function measures, for instance, the number of matching bases in the overlap.

Thus the contigs grow by greedy extension, always taking on the read that is found by following the highest-scoring overlap. The greedy algorithms can get stuck at local maxima if the contig at hand takes on reads that would have helped other contigs grow even larger. The greedy algorithms are implicit graph algorithms. They drastically simplify the graph by considering only the high-scoring edges. As an optimization, they may actually instantiate just one overlap for each read end they examine. They may also discard each overlap immediately after contig extension.

Like all assemblers, the greedy algorithms need mechanisms to avoid incorporating false-positive overlaps into contigs. Overlaps induced by repetitive sequence may score higher than overlaps induced by common position of origin. An assembler that builds on false-positive overlaps will join unrelated sequences to either side of a repeat to produce chimera.

SSAKE [80] was the first short-read assembler. It was designed for unpaired short reads of uniform length. SSAKE does not use a graph explicitly. It does use a lookup table of reads indexed by their prefixes. SSAKE iteratively searches for reads that overlap one contig end. Its candidate reads must have a prefix-to-suffix identical overlap whose length is above a threshold. SSAKE chooses carefully among multiple reads with equally long overlaps. SHARCGS [81] also operates on uniform-length, high-coverage, unpaired short reads. It adds pre- and post-processor functionality to the basic SSAKE algorithm. VCAKE [82] is another iterative extension algorithm. Unlike its predecessors, it could incorporate imperfect matches during contig extension. VCAKE was later combined with Newbler in a pipeline for Solexa+454 hybrid data [83].

Overlap/Layout/Consensus Assemblers. The OLC approach was typical of the Sanger-data assemblers. It was optimized for large genomes in software

including Celera Assembler [84], Arachne [85, 86], and CAP and PCAP [87]. The OLC approach has been reviewed elsewhere [88, 89].

Newbler [90] is widely used software distributed by 454 Life Sciences. The first release, described in the published supplement, targeted unpaired reads of approximately 100 bp as generated by the GS 20 machine. Newbler has since been revised, in particular to build scaffolds from paired-end constraints. The Newbler package offers functionality beyond de novo assembly.

The Celera Assembler [84] is a Sanger-era OLC assembler revised for 454 data [36]. The revised pipeline, called CABOG, discovers overlaps using compressed seeds. CABOG reduces homopolymer runs, that is, repeats of single letters, to single bases to overcome homopolymer run length uncertainty in data.

The de Bruijn Graph Approach. The third approach to assembly is most widely applied to the short reads from the Solexa and SOLiD platforms. It relies on ***k-mer*** graphs, whose attributes make it attractive for vast quantities of short reads. The ***k-mer*** graph does not require all-against-all overlap discovery, it does not (necessarily) store individual reads or their overlaps, and it compresses redundant sequence. Conversely, the *k-mer* graph does contain actual sequence and the graph can exhaust available memory on large genomes. Distributed memory approaches may alleviate this constraint.

The ***k-mer*** graph approach dates to an algorithm for Sanger read assembly [72] based on a proposal [74] for an even older sequencing technology; see [78] for review. The approach is commonly called a de Bruijn graph (DBG) approach or an Eulerian approach [31] based on an ideal scenario – Eulerian path assembly. Given perfect data – error-free ***k-mers*** providing full coverage and spanning every repeat – the ***k-mer*** graph would be a de Bruijn graph and it would contain an Eulerian path, that is, a path that traverses each edge exactly once. The path would be trivial to find making the assembly problem trivial by extension. Of course, ***k-mer*** graphs built from real sequencing data are more complicated.

To the extent that the data is ideal, assembly is a by-product of the graph construction. The graph construction phase proceeds quickly using a constant-time hash table lookup for the existence of each ***k-mer*** in the data stream. Although the hash table consumes extra memory, the ***k-mer*** graph itself stores each ***k-mer*** at most once, no matter how many times the K-mer occurs in the reads. In terms of computer memory, the graph is smaller than the input reads to the extent that the reads share ***k-mers***.

The Eulerian fragment assembly avoids the costly computation of pair wise alignments between reads [31]. The De Bruijn graph of a genome has as its vertices all distinct $\boldsymbol{k_1}$ tuples that occur within the sequence (where $\boldsymbol{k}$ is the word length that is used). A directed edge is inserted between $\boldsymbol{s}$ and $\boldsymbol{t}$ if there is a **k** tuple $(u_1, u_2, \ldots, u_{k-1}, u_k)$ in the genome such that $\boldsymbol{s} = (u_1, u_2, \ldots, u_{k-1})$ and $\boldsymbol{t} = (u_2, \ldots, u_{k-1}, u_k)$ if $\boldsymbol{s}$ and $\boldsymbol{t}$ appear shifted by single nucleotide. A sketch of a graph construction procedure

is shown in Fig. 2. In practice one uses the ***k***-tuples appearing in the collection of the sequence reads and a value of ***k*** between 6 and 9 or 10. In the error-free case, the genomic sequence can be read off directly as an Eulerian path through the De Bruijn graph (with repeats forming "tangles"). In real, error-prone data underrepresented ***k***-tuples, i.e. ***k***-tuples that appear less frequently than expected from the coverage rate, indicate sequencing errors and can be omitted.

Pevzner [74] explored problems that genomic repeats introduce. Repeats induce cycles in the ***k-mer*** graph. These would allow more than one possible reconstruction of the target. Idury and Waterman [72] also explored problems of real data. They added two extra types of information to the ***k-mer*** graph and named the result a sequence graph. Each edge was labeled with the reads, and positions within each read, of the sequences that induced it. Where nodes had one inbound and one outbound edge, the three elements could be compressed into one edge. This was called the elimination of singletons. Further research led to the Euler software implementation [31] for Sanger data. Impractical for large-scale Sanger sequencing projects, Euler and the DBG approach were well positioned when the Illumina platform started to produce data composed of very short unpaired reads of uniform size.

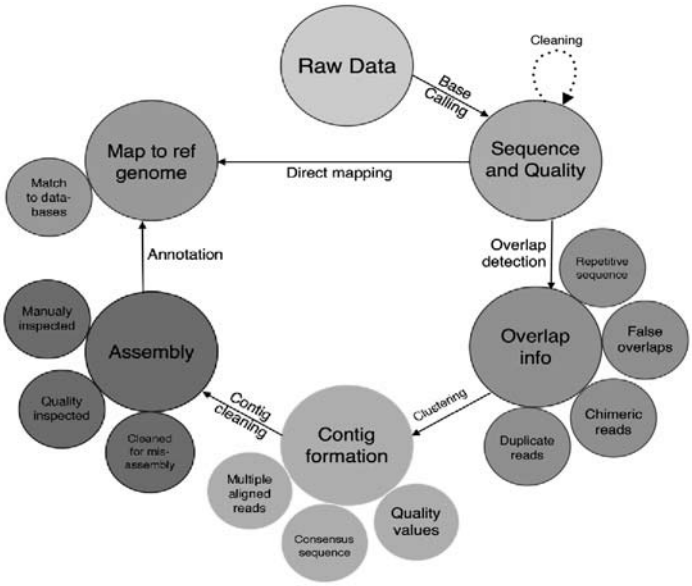


Fig.3. Assembly pipeline. A typical pipeline of a sequencing project. Sequenced reads are generated, after which they are cleaned and assembled. Following the assembly annotation and analysis can be performed. The grey line show the pipeline for massively parallel sequencing where the reads are mapped to a reference genome, while the full pipeline is for *de novo* sequencing and assembly.

6. *De novo* genome assembly

Despite a dramatic increase in the number of complete genome sequences available in public databases, the vast majority of the biological diversity in our world remains unexplored. Short Read Sequence (SRS) technologies have the potential to significantly accelerate the sequencing of new organisms.

De novo assembly of SRS data, however, requires an upgrade of the existing and the development of new software tools that can overcome the technical limitations of these technologies.

De novo genome assembly is often likened to solving a large "jigsaw puzzle" without knowing the picture we are trying to reconstruct [71]. Repetitive DNA segments correspond to similarly colored pieces in this puzzle (e.g. sky) that further complicate the reconstruction.

Mathematically, the *de novo* assembly problem is difficult irrespective of the sequencing technology used, falling in the class of ***NP-hard*** problems [28, 30, 40, 70, 78, 99], computational problems for which no efficient solution is known. Repeats are the primary source of this complexity, specifically repetitive segments longer than the length of a read. An assembler must either 'guess' (often incorrectly) the correct genome from among a large number of alternatives (a number that grows exponentially with the number of repeats in the genome) or restrict itself to assembling only the non-repetitive segments of the genome, thereby producing a fragmented assembly.

The complexity of the assembly problem has partly been overcome in Sanger projects because of the long reads produced by this technology, as well as through the use of "mate pairs" (pairs of reads whose approximate distance within the genome is known). Paired reads are particularly useful as they allow the assembler to correctly resolve repeats and to provide an ordering of the contigs along the genome.

A number of assembly programs have been developed for *de novo* assembly of NGS data. NEWBLER (Roche-applied-science.com) is distributed with 454 Life Sciences instruments and has been successfully used in the assembly of bacteria [90]. With sufficiently deep coverage, typically 25–30 times, the resulting assemblies are comparable to those obtained through Sanger sequencing [91]. Note, however, that these results do not account for the additional information provided by mate-pairs – information commonly available in Sanger data but only recently introduced to the 454 technology [92].

Some recently developed assembly tools tackle the *de novo* assembly using very short sequences (30–40 bp). SSAKE [80], VCAKE [82] and SHARCGS [81] all use a similar 'greedy' approach to genome assembly. Specifically, reads are chosen to form 'seeds' for contig formation. Each seed is extended by identifying reads that overlap it sufficiently (more than a specific length and quality cut-off)

in either the 5′ or 3′ direction. The extension process iteratively grows the contig as long as the extensions are unambiguous (i.e. there are no sequence differences between the multiple reads that overlap the end of the growing contig).

This procedure avoids mis-assemblies caused by repeats but produces very small contigs. The assembly of bacterial genomes using Illumina data created contigs that are only a few kilobases in length [28, 33, 93, 95], in contrast to hundreds of kilobases commonly achieved in Sanger-based assemblies. This fragmentation is caused in part by inherent difficulties in assembling short read data, although future improvements in assembly algorithms should overcome some limitations through more sophisticated algorithms (as was the case when Sanger sequencing was first introduced). These programs have relatively long running times, on the order of 6–10 h for bacterial assemblies [28, 93]—at least partly because of the large number of reads generated in an SRS project. By contrast, assemblers for Sanger data can assemble bacterial genomes in just a few minutes.

Another strategy for *de novo* genome sequencing uses a hybrid of NGS and Sanger sequencing to reduce costs and fill in coverage gaps caused by to cloning biases. Such an approach was followed by Goldberg *et al.* [94], who used Newbler for an initial assembly of data obtained from a 454 sequencer. They broke the Newbler contigs into overlapping Sanger-sized fragments and used Celera Assembler [27] to combine these fragments with sequence reads obtained from Sanger sequencers. This strategy proved successful in the assembly of several marine bacteria. The addition of 454 data produced better assemblies than those obtained with Sanger data alone, and for two of the genomes, the hybrid assembly enabled the reconstruction of an entire chromosome without gaps.

The assemblers named above follow the standard overlap-layout-consensus approach to genome assembly, a paradigm that treats each read as a discrete unit during the reconstruction of a genome. The previously discussed alternative algorithm proposed by Chaisson and Pevzner [95] uses a de Bruijn graph approach, an extension of the authors' prior work on assembly of Sanger data. Briefly, a de Bruijn graph assembler starts by decomposing the set of reads into a set of shorter DNA segments. A graph is constructed that contains the segments as nodes and in which two segments are connected if they are adjacent in one of the original reads. A correct reconstruction of the genome is represented as a path through this graph that traverses all the edges (an Eulerian path). By fragmenting the original reads into smaller segments, this paradigm is less affected by the short read lengths generated by SRS technologies, and it also provides a simple mechanism for combining reads of varied lengths. Chaisson and Pevzner [33, 95, 96] showed their assembler (Euler-SR) is able to generate high-quality assemblies of bacterial genomes from 454 reads and of BAC clones from Solexa reads. They also explored the use of a hybrid assembly approach (Sanger + 454) and

interestingly showed that only a small percentage of the longer reads provided information not already represented in the short reads, thus suggesting the need for a careful evaluation of the benefits of hybrid sequencing approaches.

7. Genome annotation methods for next generation sequencing technology

The highly fragmented assemblies resulting from NGS projects present several problems for genome annotation. The use of NGS technology is relatively new and few methods have been published describing how current annotation methods can be adapted to account for the various types of sequencing errors that might be present in a genome sequenced with the newer technology.

We can expect that the annotation of genomes sequenced by the new technologies will be relatively accurate for genes that are found in other species, because the primary annotation methods—sequence alignment programs—are robust in the presence of errors . Note that sequencing errors will make some of these genes appear to have in-frame stop codons, such that it might be difficult to distinguish them from pseudogenes. Nonetheless, at least the genes will be found, even if they are fragmented. By contrast, genes that are unique to an organism will be difficult to find with current annotation methods, and many of these might be missed entirely. This problem will be exacerbated by the expected small size of most contigs in assemblies of short-read sequencing projects.

The information used to annotate genomes comes from three types of analysis: (i) *ab initio* gene finding programs, which are run on the DNA sequence to predict protein-coding genes; (ii) alignments of cDNAs and expressed sequence tags (ESTs), if available, from the same or related species; and (iii) translated alignments of the DNA sequence to known proteins. These types of evidence are abundant in various amounts depending on the organism; for less well-studied species, cDNA and EST evidence is often lacking, and annotators depend much more heavily on *ab initio* prediction programs [79].

Fortunately, the main bioinformatics programs for aligning cDNAs and protein sequences to genomic DNA are robust in the presence of sequencing errors. Programs for cDNA alignment include GMAP [100], sim4 [101], Spidey [102] and Blat [103]; programs for spliced alignment of proteins to (translated) genomic DNA sequence include DPS [104] and GeneWise [105]. All of these programs must account for the fact that the target genome might not be the same strain or species as the reference cDNA or protein, so they already allow for mismatches. These sequence alignment programs should therefore work well at identifying genes even in the highly fragmented assemblies produced from short reads.

Ab initio gene finders, of which there are many, are not nearly so robust in the

presence of errors. Even with near-perfect data, the best pure *ab initio* methods for human gene finding (those not relying on alignment to other species) only identify 50–60% of exons and 20– 25% of genes correctly [106]. Gene finding in smaller eukaryotes tends to be more accurate because of their smaller introns and greater gene density, and gene finders for bacteria, archaea and viruses are very accurate, predicting >99% of protein-coding genes correctly for most genomes [107]. All of these methods assume that the DNA sequence is (mostly) correct, and certain types of errors will lead to erroneous gene predictions. In particular, any sequencing error that introduces an in-frame stop codon is likely to result in a mistaken gene prediction, because *ab initio* methods organize their searches around open reading frames [108].

8. Metagenomics and the *de novo* assembly of short sequence reads

Metagenomics is the sequencing of DNA in an environmental sample. Whereas WGS targets one genome, metagenomics usually targets several. The metagenomics assembly problemis confounded by genomic diversity and variable abundance within populations. Assembly reconstructs the most abundant sequences [109]. Simulations indicate high rates of chimera, especially in short contigs assembled from complex mixtures [110]. Studies that rely on characterization of individual reads prefer long reads [111]. The role for *de novo* genomic assembly from NGS metagenomics data should grow as NGS read lengths and NGS paired-end options increase.[28, 30, 97, 98, 119].

Most microbial species can not be cultured and even within species can be observed a huge variations in genotype (and consequently in phenotype) owing to genetic plasticity. Therefore sampling signature genes or the basic data semantics such as 16S ribosomal RNA does not give much inside into metabolic activity of a microbial community. This problem has been recently addressed by emerging field of metagenomics. Metagenomics could be regarded as a “brute force approach” whereby total DNA from microbial and/or viral population is sequenced and compared with previously sequenced genes. The high-throughput capability offered by Next Generation Sequencing methods makes them attractive for such approach.

Until now, we have considered applications that rely on mapping next-generation sequence data to available reference genome sequences. However, at least for smaller bacterial genomes, even the shortest reads can be used to effectively assemble genome sequences *de novo*, and even where complete closure of the genome is not possible, large contigs can be reliably constructed from such data provided that repeated sequences are not overly abundant. It should be noted that the continued increase in length of reads obtained by NGS platforms suggest that in the near future, *ab initio* sequencing of some eukaryotic

genomes with technologies such as Illumina or ABISOLiD is likely to be come a realistic prospect, while near-complete drafts of many microbial genomes can now be produced using the 454 technology [28, 97,98, 110, 111, 112].

The de novo assembly of sequence reads is not always necessary when comparing closely related strains: cataloguing polymorphisms relative to a reference genome is often a satisfactory goal. The main goal of both *de novo* sequencing and re-sequencing projects is to generally identify SNPs (single nucleotide polymorphisms) and other type of polymorphisms such as short insertions or deletions (collectively called indels). SNP discovery is essential for genetic mapping in eukaryotic organism with large genomes. On the other hand the detection of SNPs and rare variants in microbial genomes communities give us crucial information about the biodiversity of the prokaryote genomes and maximizes the chances of finding isolate-specific genetic differences related to important social diseases.

SNP detection and *de novo* assembly of microbial genomes are very sensible of detection of errors in the short reads obtained from the NGS technologies. The level of the errors may have a substantial impact on subsequent SNPs determination and the assembly of the genomes of the studied microbial communities.

Error detection in metagonomic NGS data. Because of the nature of metagenomic data, it is neither possible to resample the data to account for the sequencing errors that inevitably occur, nor it is possible to clearly differentiate between an error and a biological variation. Small errors in the sampled data often lead to significant changes in the results of any further analyses and studies based on the data, for example during the construction of phylogenetic trees or during the evaluation of the biological diversity in the sampled environment. For improving the quality of such studies, it is essential that an approach for detecting probable errors is devised.

There are numerous published methods for error detection and correction in NGS data, but none of them are designed to work with metagenomic data, and instead focus on applications such as de novo sequencing of genomes where the appearance of biological variations that are undistinguishable from the errors are not an issue. An example of such software is SHREC [114] which corrects errors in short-read data using a generalized suffix trie, which we use as basis for comparison.

The input data for the initial tests consists of 16S RNA short-reads. 16S RNA is very useful for metagenomic studies, because it contains highly conserved regions that can be easily isolated and compared, and at the same time it contains hypervariable regions that are greatly useful for making a distinction between different species. At the same time, this makes it more difficult to process the data. Our tests are done on sets of a few tens of thousands reads with lengths

between 300 and 500 bases. For the proposed method to be applied, the read sets need to be filtered of obvious noise and then aligned to each other so that all functional parts can be easily compared to one another [113].

The basic idea behind error correction is that if a given a bit of data such as a single base appears too little in the dataset it is more likely for it to be an error than a biological variation, and a threshold can be established using the error rate of the sequencing equipment. This can be significantly improved if during the assessment, a discrimination on similarity is made by giving higher weights to reads that are similar to the read in question in the evaluated region.

8.1. Algorithms and software solutions for *de novo* sequence assembly and annotation in metagenomes

The metagenomics assembly problem is confounded by genomic diversity and variable abundance within populations. Assembly reconstructs the most abundant sequences [115]. Simulations indicate high rates of chimera, especially in short contigs assembled from complex mixtures [28].Studies that rely on characterization of individual reads prefer long reads [30]. The role for *de novo* genomic assembly from NGS metagenomics data grows as NGS read lengths and NGS paired-end options increase.

Currently, metagenomics projects are focusing on bacterial species, which simplifies the annotation problem somewhat. Because bacterial genomes are gene-rich, a large majority of sequence reads should contain fragments of protein-coding genes. However, the usual annotation approach to bacterial genomes, which relies on (highly accurate) bacterial gene finders, does not work for environmental mixtures. Annotators have thus far relied on a simple but effective BLAST-based strategy: for each read, they use *tblastn* [103] to translate the sequence in all six frames and search a protein database for matches. For example, this annotation strategy was used for the Sargasso Sea project [109], which sampled the bacterial population of a region of the Atlantic ocean. A BLAST-based strategy can also work for short reads, although accuracy declines as reads and contigs get shorter.

Huson *et al.* [116] have developed a method – MEGAN –to enhance this translated BLAST strategy, making it more robust with short-read sequencing data. Rather than providing a detailed annotation of genes, MEGAN attempts to characterize the phylogenetic makeup of a sample, which often is the primary goal of a metagenomics sequencing project. In other words, the goal is to identify the species present rather than the precise identities and locations of all the genes in a mixed sample.

In another recent development, Krause *et al.* [117] enhanced a translated BLAST approach in an effort to make it more robust to the sequencing errors

common in SRS projects. Their CARMA systems combines translated BLAST searches with a postprocessing step that merges protein fragments across frameshifts. They tested their system on a synthetic metagenomic dataset sequenced with 454 technology and were able to identify many of the frameshifts and in-frame stop codons caused by sequencing errors. However, accuracy was substantially lower than a standard bacterial gene finder would obtain on a genome assembled from Sanger sequencing data.

The MetaGene system of Noguchi *et al.* [118] is designed specifically for metagenomic data from short reads. It uses two dicodon frequency tables, one for bacteria and another for archaea, and applies them based on the GC-content of the sequence fragment. It could reproduce >90% of the gene annotation from the Sargasso Sea project using the contigs generated from that project (which were ~1 kbp in length on average).

9. Parallel computing opportunities

The processing of a great number of large samples of metagenomic data creates a real computational challenge. While it is interesting to discuss parallel execution of the error detection algorithm itself, the pre-processing step is both more expensive and a more challenging problem to solve. Due to an inherent similarity between the preprocessing and some of the further studies, such as reconstruction of phylogenetic trees, the usefulness of parallel computation would double. [120, 121, 122, 123, 124].

The popular algorithms for multiple sequence alignment are based on the construction of a rough similarity score matrix for the reads which is then used to construct a tree in which similar reads are neighbours using the neighbour joining method or some derivative, then neighbours are aligned against each other. While it is trivial to compute the similarity matrix in parallel, the same is difficult for the construction of the tree.

The direct approach requires the use of shared memory which is neither popular in computer clusters, nor does it scale well. To overcome the need for shared memory, there are published methods that rely on heuristics to split the data into groups by similarity, then every group can be aligned on its own and all the groups can be aligned in parallel. This is similar to the approach that we're currently using in our error detection pipeline, although the final choice for a suitable heuristic has not been made yet.

The error evaluation is almost independent for each region, so it won't be a challenge to run multiple regions in parallel, and the slight window overlap would waste an insignificant amount of memory. However, it is one of the less intensive task in the pipeline. If the number of reads is n, and the maximum

read length is m, then the matrix calculation and the error evaluation are of time complexity $O(mn^2)$, while the tree construction is of time complexity $O(n^3)$, and thus it is the main focus, although some of the heuristics for splitting the dataset can reduce the time complexity or otherwise significantly reduce the CPU time required for the tree construction, so a parallel implementation for all the steps is preferable.

Parallel processing would be crucial for the actual studies as well. Essentially, the construction of a phylogenetic tree is almost identical to the algorithm for multiple sequence alignment. The only difference is that the similarity matrix is computed over sequences that have been already aligned and a different metric for the distance is used, so solving either would also solve the other one, and both steps have essentially the same computational requirements. Any differences created by the changing of the choice of distance or the sequence gaps are negligible.

The data sets for a given scientific experiment are not only big, but they also tend to be numerous. In fact they tend to grow more in number than in size, which makes them an excellent target for parallel processing regardless of the algorithms used during the data processing, and regardless of the nature of the task that is to be solved.

10. Computational Infrastructure Challenges with Next Generation Sequencing

The success of next-generation sequencing in biology is unquestioned. However, access to these technologies is not without challenges at multiple levels. The discussion above focused primarily on technical challenges encountered in generating the sequence data, all which are being overcome at a relatively fast pace. However, the ability to generate large sequence datasets from an unlimited number of plant species can cause data overload at laboratory, institutional and community scale. Indeed, the infrastructure costs for data storage, processing and handling are typically more than the costs of generating the sequence. Coupled with the promise of even more throughput in the coming years via the third-generation sequencing platforms, data storage and handling issues will continue to grow. Major changes in how computation is performed on large genome datasets have occurred over the last decade. In the Sanger sequencing era, genome assembly was usually performed at a genome center using a computational infrastructure called a compute cluster that required extensive investment in CPUs and data storage. However, in the last decade, there have been numerous advances in computer technology and bioinformatic tools that now put access to compute resources within the fiscal and practical reach of a single investigator. Below, we aim to provide a brief explanation of new computing approaches that have

the potential to 'democratize' genome data analysis in the same way that the next-generation sequencing platforms have enabled a wide range of biologists to design and generate their own sequence datasets.

10.1. Compute and grid clusters

In a compute cluster, the computers are networked to each other using fast networking technologies such gigabit Ethernet or Infiniband. Storage, a critical issue with genome sequence datasets and the associated bioinformatic data analyses, is local to the compute cluster. Redundant disk arrays are used for optimal data access performance and to maintain data integrity. Jobs are submitted and managed on the compute cluster using software such Sun Grid Engine (now Oracle Grid Engine, http://wikis.sun.com/ display/GridEngine/ Home) or Condor (http://www.cs.wisc.edu/condor/). High-speed networking enables parallel computing solutions that distribute the jobs across the nodes of the cluster. Several genomics applications now support use of parallel environments, including genome assemblers such as ABySS [125] and Velvet [126], genome annotation pipelines such as MAKER [127], and sequence aligners such as mpiBLAST (http://www.mpiblast.org/). The power, cooling and system administration requirements for running a compute cluster can place it out of reach for a laboratory or small department. However, access to a compute cluster that is maintained by a dedicated staff is available at most institutions. For groups that wish to run their own compute cluster, cluster management software is available, such as the Rocks cluster distribution (http://www.rocksclusters.org/), which automates many of the tasks related to systems administration of the compute cluster.

Since the early 2000s, grid computing infrastructures, such as National Lambda Rail (http://www.nlr.net/), have become available. In grid computing, the computational resources are geographically distributed and linked together using the Internet or next-generation research networks. Grid computing resources available to researchers include the TeraGrid [128] and the Open Science Grid (http://www.opensciencegrid.org/). Some disadvantages of grid computing for genome informatics include uploading large datasets to the grid (such as databases and sequence read files), installing genomics programs on the grid, and connectivity issues. Data access and storage on the grid can also be complex due to the heterogeneous data storage resources and file systems available at each grid site, slow data transfer speeds between grid sites, limited storage space quotas, and limited or non-existent data back-up options.

10.2. Cloud computing

Cloud computing has recently come to the forefront of highperformance

computing infrastructure for bioinformatics [129]. Using a technology called virtualization, the cloud computing infrastructure allows the user to create a virtual compute cluster through 'instances' of virtual machines on host servers. A virtual machine is a file that contains an operating system, programs and possibly data that can be 'booted' on a host and run as an independent computer. Similar functionality for desktop computer users is provided by programs such as VMWare (http://www.vmware.com) and Virtualbox (https://www.virtualbox.org/). A popular cloud infrastructure platform is the Elastic Compute Cloud provided by Amazon (https://aws.amazon.com/ec2/). The EC2 offers several tiers of virtual machine 'instances' at several price points, from micro instances to compute cluster instances. Storage is provided by a Simple Storage Service (S3) or by attaching an Elastic Block Store (EBS) to the virtual machine instance. EBS storage is not persistent, and is suited for storing temporary results and other local storage needs. In contrast, S3 storage is persistent but attracts a fee. Although importing data is currently free, users are charged for exporting data from the EC2.

Cloud infrastructure software that emulates the Amazon EC2 platform, such as Eucalyptus (http://open.eucalyptus.com/), has been developed, thereby allowing creation of private clouds. This has the benefits of running custom virtual machine images, dynamically scaling the cluster size, and retaining control over security and data storage. The compatibility of the cloud infrastructure and virtual machine images also allows for hybrid clouds, i.e. private and commercial/community-based, in the event that additional computing power is needed. Means to make cloud computing infrastructure available to researchers include the FutureGrid (https://portal.futuregrid.org/), a next-generation extension to the TeraGrid.

The creation of virtual machine images requires programming or system administration knowledge. To alleviate this bottleneck, several virtual machine images for genome informatics are available. CloVR contains an entire bacterial assembly/annotation pipeline and a metagenomic sequencing pipeline [130], whereas CloudBioLinux (http://cloudbiolinux.org/) contains a number of genome sequencing, assembly and annotation programs. Utilizing on these prebuilt machine images are integrated systems such as Galaxy CloudMan [131] that can create an entire compute cluster in the cloud, allowing users with no computational experience to run a bioinformatics pipeline. The iPlant Collaborative provides computational infrastucture to support plant biology (http://www.iplantcollaborative.org/), including Atmosphere, a cloud infrastructure developed to support computational efforts in plant biology.

The cloud, when combined with frameworks that distribute computational tasks across many nodes, allows new approaches to processing the large amounts of data produced by next-generation sequencing. Early examples of this approach include Cloudburst, an aligner used to map next-generation reads [132], Crossbow,

a pipeline used to map next-generation sequencing reads onto a reference genome and call single nucleotide polymorphisms [133], Myrna, which is a differential gene expression pipeline [134], and Contrail, a distributed genome assembly program for nextgeneration reads (http://contrail-bio.sf.net).

11.Discussion and conclusions

New high-throughput sequencing technologies have emerged. However, the sequencing methods as well as the computational tools and infrastructure have to be further improved, to allow new approaches for complete *de novo* assembly for large genomes with these technologies. However, for now there are not too much projects available for using of parallel computing methods based on concurrent computational infrastructure.

The crucial problem with *de novo* assembly algorithms demanding more computing resources and high-throughput infrastructure requires some new approaches for finding the optimal solution. The short read lengths of some of the major technologies seem to be a momentous disadvantage and the high number of reads produced might not be able to compensate for this handicap. However, all manufacturers aim to increase the read lengths. Currently, areas on able approach to the assembly of such short sequences could include data from low coverage Sanger sequencing aside with data from the NGS platforms. Although hybrid data set approaches are cumbersome [94,95, 96], they have already been shown to produce useful assemblies [97,98].

A consequence of the dissemination of genome sequencing infrastructure to a wider group of scientists is that there is increased breadth and depth in the biological questions being addressed by genomics. Thus to meet the unprecedented flood of data generated by the next generation of DNA sequencers, bioinformatics groups working with NGS data found it necessary to respond quickly and efficiently to the informatics and infrastructure demands. Centralized facilities newly facing this challenge need to anticipate time and design considerations of necessary components, including infrastructure upgrades, staffing, and tools for data analyses and management.

With infrastructure costs decreasing substantially, such that individual labs can equip themselves with their own sequencing instrument, there will be further innovation in genomic questions, applications and approaches. However, problems with data handling and data volume are already evident. Certainly, new computing and bioinformatic approaches will advance; however, the sheer amount of data storage and handling required for the ever-growing output of sequencing platforms will continue to be a major challenge for all genomicists and IT experts. However, given the continuing advances in information technology, these technical challenges will be overcome, and we can focus on

wider interdisciplinary scientific domains such as system biology, to complete our quest for complete knowledge of the biology of studied organisms.

12. References

[1] Schuster S.C. (2007) Next-generation sequencing transforms today's biology. Nat. Methods, 5, 16–18.

[2] Qin J. et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464, 59–65.

[3] Van Tassell C.P. et al. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods, 5, 247–252.

[4] Alkan C. et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat. Genet., 41, 1061–1067.

[5] Medvedev,P. et al. (2009) Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods, 6, S13–S20.

[6] Taylor K.H. et al. (2007) Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. Cancer Res., 67:8511–8518.

[7] Sultan M. et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science , 321, 956–960.

[8] Guffanti A. et al. (2009) A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics, 10, 163–179.

[9] Auffray C. et al. (2009) Systems medicine: the future of medical genomics and healthcare. Genome Med., 1, 1–2.

[10] Miller J.R. et al. (2010) Assembly algorithms for next-generation sequencing data. Genomics, 95, 315–327.

[11] Horner D.S. et al. (2010) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief. Bioinformatics, 11, 181–197.

[12] Ruffalo, M., LaFramboise T., Koyuturk M. (2011) Comparative analysis for next generation sequence alignmen. Bioinformatics 27(20):2790-2796.

[13] Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinformatics, 11, 473–483.

[14] Langmead,B. et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10, R25.

[15] Li,R. et al. (2008b) SOAP: short oligonucleotide alignment program. Bioinformatics, 24, 713–714.

[16] Li,R. et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 25, 1966–1967.

[17] Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 589–595.

[18] Hach,F. et al. (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat. Methods, 7, 576–577.

[19] Novocraft (2010) http://www.novocraft.com/.

[20] Rumble,S.M. et al. (2009) Shrimp: accurate mapping of short color-space reads. PloS Comput. Biol., 5, e1000386.

[21] Droege M, Hill B. (2008) The Genome Sequencer FLX System—longer reads, more applications, straight forward bioinformatics and more complete data sets. J Biotechnol 136:3–10.

[22] Bennett S. (2004) Solexa Ltd. Pharmacogenomics, 5:433–8.

[23] Porreca GJ, Shendure J, Church GM. (2006) Polony DNA sequencing. Curr ProtocMol Biol Chapter 7:Unit:7–8.
[24] Harris TD, Buzby PR, et al. (2008) Single-molecule DNA sequencing of a viral genome. Science 320:106–9.
[25] Branton D, Deamer DW, et al. (2008) The potential and challenges of nanopore sequencing. NatBiotechnol 26:1146–1153.
[26] Pettersson E, Lundeberg J, Ahmadian A. (2009) Generations of sequencing technologies. Genomics 93:105–111.
[27] Myers E, Sutton G, et al., (2000):**A whole-genome assembly of Drosophila.** *Science, 2***87**(5461)**:**2196.
[28]Miller J., Koren S., Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315-327
[29] Scheibye-Alsing K., et al. (2009) Sequence assembly. Computational Biology and Chemistry 33:121-136
[30] Koren S., et al. (2012) Hybrid error correction and de novo assembly pf single-molecule sequencing reads. Nature Biotechnology 30,693–700
[31] Pevzner P.A., Tang H., Waterman (2001) An Eulerian path approach to DNA fragment assembly. PNAS 98:9748-9753
[32] Compeau P.E.C., Pevzner P.A., Tesler G. (2011) How to apply de Bruijn graphs to genome assembly. Nature Biotechnology 29(11):987-991
[33] Chaisson M., Brinza D., Pevzner P.A. (2009) De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res 19(2): 336–346.
[34]Simpson J.T., Kim Wong, Shaun D. Jackman S.D, et al. (2009) ABySS: A parallel assembler for short read sequence data. Genome Res. 19:1117-1123
[35] Ruiqiang Li, Hongmei Zhu, Jue Ruan, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing Genome Res. 20: 265-272
[36] Miller J, Delcher A, et al. (2008) A**ggressive assembly of pyrosequencing reads with mates.** *Bioinformatics* 24**:**2818-2824.
[37] Dalca AV, Brudno M. (2010) Genome variation discovery with high-throughput sequencing data. Brief Bioinform 11:3–14.
[38] Pepke S, Wold B, Mortazavi A. (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6:S22–32.
[39] Cokus SJ, Feng S, Zhang X, et al. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452:215–219.
[40] Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat Methods 6:S6–12.
[41] Malhis N, Butterfield YSN, Ester M, et al. (2009) Slider–maximum use of probability information for alignment of short sequence reads and SNP detection. Bioinformatics 25: 6–13.
[42] Malhis N, Jones SJ. (2010) High quality SNP calling using Illumina data at shallow coverage. Bioinformatics 26:1029–1035.
[43] Altschul SF, Gish W, Miller W, et al. (1990) Basic local alignment search tool. J Mol Biol 215:403–410.
[44] Altschul SF, Madden TL, Schaffer AA, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.
[45] Smith TF, Waterman MS. (1981) Identification of common molecular subsequences. J Mol Biol., 147:195–197.
[46] Gotoh O. (1982) An improved algorithm for matching biological sequences. J Mol Biol., 162:705–708.

[47] Ma B, Tromp J, Li M. (2002) PatternHunter: faster and more sensitive homology search. Bioinformatics 18:440–445.
[48] Li R, Li Y, Kristiansen K, etal. (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714.
[49] Schatz M. (2009) CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25:1363–1369.
[50] Chen Y, Souaiaia T, Chen T. (2009) PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. Bioinformatics 25:2514–2521.
[51] Clement NL, Snell Q, Clement MJ, et al. (2010) The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides tides from next-generation sequencing. Bioinformatics 26:38–45.
[52] Jiang H, Wong WH. (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 24:
2395–2396.
[53] Rumble SM, Lacroute P, Dalca AV, et al. (2009) SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol. 5:e1000386.
[54] Weese D, Emde AK, Rausch T, et al. (2009) RazerS fast read mapping with sensitivity control. Genome Res 19:1646–1654.
[55] Farrar M. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics 23:156–161.
[56] Homer N, Merriman B, Nelson SF. (2009) BFAST: an alignment tool for large scale genome re-sequencing. PLoS One 4:e7767.
[57] Eppstein D, Galil Z, Giancarlo R, Italiano GF. (1990) Sparse dynamic programming. In: SODA. Philadelphia: Society for Industrial and Applied Mathematics, 513–522.
[58] Slater GSC, Birney E. (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.
[59] Myers EW. An O(ND) Difference algorithm and its variations. Algorithmica 1986;1(2):251–66.
[60]Abouelhoda MI, Kurtz S, Ohlebusch E. (2004) Replacing suffix trees with enhanced suffix arrays. J DiscreteAlgorithms 2:53–86.
[61] Ferragina P., Manzini G. Opportunistic data structures with applications. (2000) In: Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS 2000), Redondo Beach, CA, USA. 2000;390–398.
[62] Phillippy A.M., Schatz M.C., Pop M., (2008) Genome assembly forensics: finding the elusive mis-assembly, Genome Biol. 9 R55.
[63] Green Laboratory (1994) Phred, Phrap, Consed Documentation. http://www.phrap.org/phredphrapconsed.html .
[64] Huang X., Madan A. (1999) CAP3: A DNA sequence assembly program. Genome Res. 9(9):868-877.
[65] Gilchrist M., Zorn A., Voigt J., Smith J., Papalopoulu N., Amaya E. (2004) Defining a large set of full-length clones from Xenopus tropicalis EST project. Dev.Biol. 271(2):498-516.
[66] International Human Genome Sequencing Consortium (2004) Finishing the euchromatin sequence of the human genome. Nature 431 (October (7011)):931-945.
[67] Chaison M., Pevzner P. (2008) Short read fragment assembly of bacterial genomes. Genome Res. 18(2):324-330. http://genome.cshlp.org/cgi/content/abstract/18/2/324 .
[68] Hernandez D., et al. (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res. 18(5):802-809.
[69] Kececioglu J., Ju J. (2001) Separating repeats in DNA sequence assembly. In proc: Annual Conference on Research in Computational Molecular Biology, pp. 176-183.
[70] Li Z., Chen Y., et al (2012) Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-brujin-graph. Briefings in Functional Genomics 11(1):25-37.

[71] Schatz M.C., Delcher A.L., Salzberg S.L., (2010) Assembly of large genomes using second-generation sequencing. Genome Res. 20:1165-1173.
[72] Idury R.M., Waterman M.S. (1995) A new algorithm for DNA sequence assembly, J Comput Biol 2:291-306.
[73] De Bruijn, N. G. (1946). A Combinatorial Problem. Koninklijke Nederlandse Akademie v. Wetenschappen 49: 758–764.
[74] Pevzner P.A. (1989) L-tuple DNA sequencing: computer analysis. J. Biomol. Struct. Dyn. 7:63-73.
[75] Pevzner P.A., Tang H., Tesler C. (2004) De novo repeat classification and fragment assembly, Genome Res. 14:1786-1796.
[76] Zhi D., Raphael B.J., Price A.L., Tang H., Pevzner P.A. (2006) Identifying repeat domains in large genomes. Genome Biol.7:R7.
[77] Nagarajan N., Pop M. (2009) Parametric complexity of sequence assembly: theory and applications to next generation sequencing, J. Comput. Biol. 16:897–908.
[78] Pop M. (2009) Genome assembly reborn: recent computational challenges, Brief. Bioinform. 10:354–366.
[79] Pop M., Salzberg S. (2008) Bioinformatics challenges of new sequencing technology. Trends Genet. 24:142–149.
[80] Warren R.L., Sutton G.G., Jones S.J., Holt R.A. (2007) Assembling millions of short DNA sequences using SSAKE, Bioinformatics 23:500–501.
[81] Dohm J.C., Lottaz C., Borodina T., Himmelbauer H. (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing, Genome Res. 17:1697–1706.
[82] Jeck W.R., Reinhardt J.A. et al. (2007) Extending assembly of short DNA sequences to handle error, Bioinformatics 23:2942–2944.
[83] Reinhardt J.A., Baltrus D.A. et al. (2009) De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae, Genome Res. 19:294–305.
[84] Myers E.W., Sutton G.G. et al. (2000) A whole-genome assembly of Drosophila. Science 287 (2000) 2196–2204.
[85] Batzoglou S., Jaffe D.B., et al., (2002) ARACHNE: a whole-genome shotgun assembler, Genome Res. 12 177–189.
[86] D.B. Jaffe, J. Butler, et al. (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2, Genome Res. 13:91–96.
[87] Huang X., Yang S.P., (2005) Generating a genome assembly with PCAP. Curr. Protoc. Bioinformatics, Chapter 11 (2005) Unit11 3.
[88] Batzoglou S., (2005) Algorithmic Challenges in Mammalian Genome Sequence Assembly, in: M. Dunn, L. Jorde, P. Little, S. Subramaniam (Eds.), Encyclopedia of genomics, proteomics and bioinformatics, John Wiley and Sons, Hoboken (New Jersey).
[89] Pop M. (2005) DNA sequence assembly algorithms, in: McGraw-Hill (Ed.), McGraw-Hill 2006 Yearbook of Science and Technology, McGraw-Hill, New York, 2005.
[90] Margulies M., Egholm M. et al., (2005) Genome sequencing in micro fabricated high-density picolitre reactors, Nature 437:376–380.
[91] Sanger F., Nicklen S., Coulson A.R., (1977) DNA sequencing with chain-terminating inhibitors, PNAS, 74:5463–5467.
[92] Khouri H., Kravitz S.A., et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes, PNAS 103:11240–11245.
[93] Diguistini S., Liao N.Y., et al. (2009) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data, Genome Biol. 10:R94.

[94] Goldberg S.M.D., et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. PNAS USA 103:11240–11245.
[95] Chaisson, M.J., et al. 2008. Short read fragment assembly of bacterial genomes. Genome Res. 18, 324–330
[96] Chaisson M., Pevzner P., Tang H. (2004) Fragment assembly with short reads, Bioinformatics 20:2067–2074.
[97] Boisvert S., Laviolette F., Corbeil J. (2010) Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comp Biology 17(11): 1519–1533.
[98] Bashir A., Klammer A.A., et al. (2012) A hybrid approach for the automated finishing of bacterial genomes. Nature Biotechnology 30(7):701-707.
[99] Medvedev P, et al. Computability and equivalence of models for sequence assembly. Lecture Notes Comput Sci 2007;4645:289–301.
[100] Wu TD, Watanabe CK. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 2005;21:1859–1875.
[101 Florea L, et al. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8:967–974.
[102] Wheelan SJ, et al. (2001) Spidey: a tool for mRNA-to-genomic alignments. Genome Res 11:1952–1957.
[103] Kent W.J. (2002) BLAT – the BLAST-like alignment tool. Genome Res., 12:656–664.
[104] Huang X, et al. (1997) A tool for analyzing and annotating genomic sequences. Genomics 46:37–45.
[105] Birney E, et al. (2004) GeneWise and Genomewise. Genome Res 14:988–995.
[106] Guigo R, et al. (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol 2006;7(Suppl 1):S21–S31
[107] Sommer D., et al. (2007) Minimus: a fast, lightweight genome assembler. BMC Bioinform 8:64.
[108] Conesa A., Gotz S., et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674-3676.
[109] Rusch D.B., Halpern A.L., et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific, PLoS Biol. 5.e77.
[110] Mavromatis K., Ivanova N., et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods, Nat. Methods 4:495–500.
[111] Wommack K.E., Bhavsar J., Ravel J. (2008) Metagenomics: read length matters, Appl. Environ. Microbiol. 74:1453–1463.
[112] MacLean D., Jones J.D.G., Studholme D.J. (2009) Application of "next-generation" sequencing technologies to microbial genetics. Nature Reviews Microbiology 7(4):287-296.
[113] Krachunov M., Vassilev D. (2012) An approach to a metagenomic data processing workflow. In book of abstracts of BIOCOMP 2012, September 19-21.
[114] Schroder J et al. (2010) SHREC: A short read error correction method, Bioinformatics, 25(17):2157-2163
[115] Thomas et al. (2012) Metagenomics – a guide from sampling to data analysis. Microbial informatics and experimentation 2:3.
[116] Huson, D.H. et al. (2007) MEGAN analysis of metagenomic data, Genome Res, 17, 377-386
[117] Krause, L. et al. (2008) Phylogenetic classification of short environmental DNA fragments, Nucleic Acids Res, 36, 2230-2239.
[118] Noguchi,H.,Park,J.andTakagi,T.(2006)MetaGene: prokaryotic gene finding from environmental genomes shotgun sequences. NucleicAcidsRes., 34,5623-5630.
[119] Kunin,V.,Copeland,A.,Lapidus,A.,Mavromatis,K.and Hugenholtz,P. (2008) A bioinformatician's guide to metagenomics. Microbiol.Mol.Biol.Rev., 72:557-578.

[120] Janaki, C. and Joshi, R.R. (2003) Accelerating comparative genomics using parallel computing. In Silico Biology., 3:429-440.
[121] Kleinjung, J., Douglas, N. and Heringa, J. (2002). Parallelized multiple alignment. Bioinformatics., 18, 1270-1271.
[122] Augen, J. (2003). In silico biology and clustered supercomputing: shaping the future of the IT industry. Biosilico., 1:47-49.
[123] Bader, D. A. (2004). Computational Biology and High-Performance Communications of the ACM., 47, 35-40.
[124] Jackson B,G., Schnable P. and Aluru S.(2009) Parallel short sequence assembly of transcriptomes. BMC Bioinformatics, 10(Suppl 1):S14
[125] Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123.
[126] Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.
[127] Cantarel, B.L., Korf, I., et al. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188–196.
[128] Beckman, P.H. (2005) Building the TeraGrid. Philos. Transact. A Math. Phys. Eng. Sci. 363:1715–1728.
[129] Stein, L.D. (2010) The case for cloud computing in genome informatics. Genome Biol. 11, 207.
[130] Angiuoli, S.V., et al. (2011) CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics 12, 356.
[131] Afgan, E., et al. (2010) Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics 11(Suppl. 12), S4.
[132] Schatz, M.C. (2009) CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25, 1363–1369.
[133] Langmead, B., Schatz, M.C., Lin, J., Pop, M. and Salzberg, S.L. (2009) Searching for SNPs with cloud computing. Genome Biol. 10, R134.
[134] Langmead, B., Hansen, K.D. and Leek, J.T. (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol. 11, R83.

Requirements for Cloud Service Discovery Systems

Georgi Pashov, Kalinka Kaloyanova

Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
gdpashov@abv.bg, kkaloyanova@fmi.uni-sofia.bg

Abstract. Availability of efficient Cloud service discovery techniques is a key factor for a broader adoption of Cloud computing. Defining business and architectural constraints is the core element of this process. In this paper, we propose a systematic approach for defining requirements for Cloud service discovery systems. Our approach is based on a strong analysis of the system's characteristics which is used to determine the most essential functional and non-functional requirements. A taxonomy of Cloud service discovery systems is designed in order to develop an appropriate architectural model.

Keywords: Cloud services, Cloud service discovery, Requirements for Cloud Service Systems.

1 Introduction

After more than a decade of use, adoption of Cloud computing continues to grow and evolve within the enterprise application markets. More and more enterprises are investigating the possibilities of using Cloud services in their business processes [3]. Despite the significant progress there are still a few challenges to a broader adoption of Cloud computing. Efficient service discovery is one of them.

Service discovery is an automated process of identifying services that fulfil the consumer's requirements. There is not yet an automated Cloud services discovery engine in operation although there are a few researches on the topic. Currently consumers have to search for Cloud services manually using conventional web search engines as Google, Yahoo, etc. and then to visit the found web pages one by one in order to investigate them in details. Manual searching has some drawbacks, e. g. the most appropriate services may not be found amongst the many web pages, it is a time consuming job, etc. That may seriously compromise the usage of Cloud computing. Manual searching is becoming even more challenging as the volumes and varieties of Cloud services are constantly growing. All this indicates that efficient Cloud service discovery techniques have to be established. Although there are many discovery techniques in Grid and other distributed systems, they cannot be transferred directly to the Cloud without taking into account the specifics of searching in that particular environment.

V. Dimitrov (Editor): ISGT'2012. ISSN 1314-4855
Proceedings of the 6th International Conference on
Information Systems and Grid Technologies, Sofia, June 1-3., 2012.

The purpose of this article is to review the existing approaches for Cloud service discovery and to outline the requirements for such systems. The rest of this paper is organized as follows. In the next section we overview the existing approaches for Cloud service discovery. In section 3 we present our multi-criteria decision making approach for developing an adequate Cloud service discovery mechanism. In section 4 we present a taxonomy designed by us in order to determine the appropriate architectural models. The last section contains conclusion and future work.

2 Overview of the Existing Approaches for Cloud Service Discovery

Although Cloud service discovery is a relatively new topic there are already a few approaches proposed on that subject.

In [8] Ranjan et al. present a layered Peer-To-Peer (P2P) Cloud provisioning architecture whose main goal is to unify the processes of provisioning and usage of Cloud services. Service discovery (as well as monitoring, fault management, scheduling and other supporting services) is part of the core services layer which is regarded as Platform as a Service. The proposed architecture provides decentralized service discovery and load-balancing between cloud components based on Distributed Hash Table (DHT) over P2P network (structured P2P).

The P2P paradigm structured as DHT is suggested in [9] for a general resource information system (i.e. including infrastructure components presented as services), which maintains dynamic data about resources in a Cloud datacenter. It overcomes many limitations of existing Grids solutions taking advantage of the Cloud-specific context. This system forms a P2P cluster of dedicated super-peers, where datacenter resource information is structured as DHT with a non-traditional keyspace partitioning algorithm that trades off better performance and fault-tolerance capabilities for disadvantages that are not of importance to the Cloud. In order to improve the performance of the partitioning algorithm it is made context-driven [10]. Usually, for resource data persistency, Cloud datacenters do not restrict all the data to be kept in a certain type of storage. In [10] instead, the authors propose that different types of resources can choose an optimal storage for their data with respect to the specific resource semantic requirements. Although data is spread into different locations and fetched via different types of queries, the proposed algorithm still implements efficient filtering for cross-resource-type searches. This is achieved by employing the techniques of logical decomposition of Boolean expressions and executing partial filters against the responsible parties.

Zhou et al. propose a hybrid P2P approach [11] to service discovery in the Cloud which implements unstructured P2P paradigm and combines various techniques in order to increase the efficiency of searching – techniques such as one-hop replication, semantic aware message routing, topology reorganization,

and supernodes. The proposed discovery system is hosted over an unstructured P2P network which is voluntarily formed by service consumers. Upon arrival, each node exchanges its information about the Cloud services it hosts with its neighbouring nodes (one-hop replication). Each node maintains local kd-tree index (no global knowledge is required) for description of all services which it is aware of. Thanks to that, both a point query and a range query can be similarly handled. If a node becomes "knowledgeable enough" in the services that other neighbours host, that is, the number of its neighbours exceeds a threshold, and considers itself capable of dealing with a large number of incoming queries, the node can elect itself a supernode. The system employs semantic-aware routing protocols and topology reconstruction techniques: When forwarding a service request, the node decides the next-hop node for the request message in response to the semantic similarity between the service request and the service information of its neighbours. Moreover, a node proactively seeks new neighbours that would satisfy the service requests in the future with high probability. As a result, the P2P network topology is reformulated.

In [4] Han et al. present an ontology-enhanced Cloud service discovery system (CSDS) that automates searching of Cloud services over the Internet. Initially conventional search engines as Google, Yahoo, etc. are used to locate web-pages of potential interest. Then the result set is filtered in order to find the most relevant web-pages using various heuristics as counting the frequencies of evidence phrases for example. On the next step the system consults Cloud ontology to determine the similarity between available and searched services using three reasoning methods: similarity, equivalent and numerical reasoning. The web-pages are ranked according to the calculated similarity and web-pages with the highest service utility are selected as the best services for the user.

Similar approaches which exploit semantic similarity reasoning for Cloud service discovery are discussed in [1] and [6].

3 Multi-criteria approach for Cloud Service Discovery Systems' Requirements

Our approach for developing an adequate Cloud service discovery mechanism is based on a multi-criteria decision making. Several elements that help for a reasonable selection of core functional capabilities, quality constraints and architectural model are discussed. This approach includes several steps: (1) discussing the generic architectural model; (2) discovering the usage patterns and processes in which CSDSs are involved; (3) defining the most essential functional requirements; (4) defining the most critical non-functional requirements; and (5) determining the architectural model.

3.1 Generic Architectural Model

The most appropriate architectural model for such kind of systems, with many and various providers and consumers, is the service broker model in which a service broker acts as an intermediary agent between providers and consumers. This approach allows broad, easy and cost efficient way for matchmaking consumers' needs to providers' offers.

All service providers will require subscription to such a mechanism, along with all the necessary information such as functional capabilities, quality constraints and interfaces, in order to advertise their services to the consumers. The CSDS makes these services available to the potential consumers, enterprises or individuals, by matchmaking consumers' queries to available services. A high-level overview of CSDS's generic architecture is shown in fig. 1.

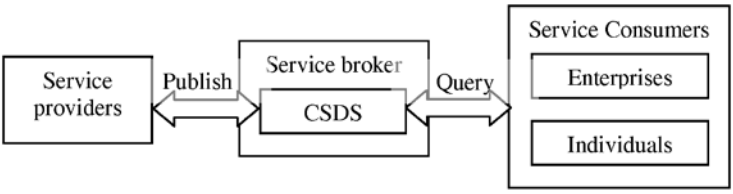


Fig. 1. Generic architecture of Cloud service discovery system.

3.2 Usage Patterns for Cloud Service Discovery Systems

CSDSs are used mainly in the initial phases of the Cloud service usage's lifecycle when the decision about using Cloud services is taken (fig. 2). Regarding that the usage of CSDS is business critical for choosing the services that match best to the consumers' needs.

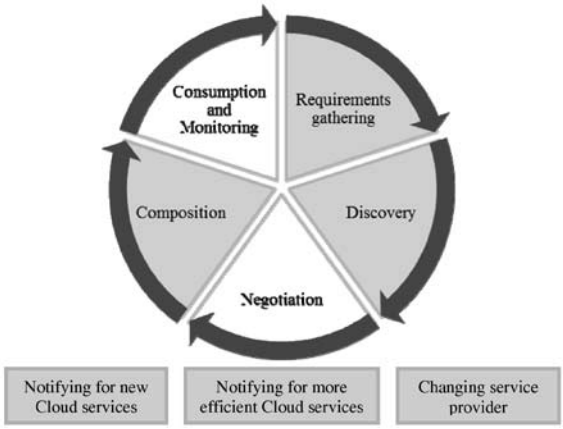


Fig. 2. Usage patterns for Cloud service discovery systems.

The first phase in the process of choosing a suitable Cloud service is requirements gathering [5]. During this phase the functional requirements, non-functional requirements (as performance, service availability, etc.) and interfaces are defined. The complete and accurate requirement analysis at this phase guarantees that the corresponding Cloud services will reach the required standards. As requirements gathering itself is a challenging process, we consider that the CSDSs could be used in requirement analysis as a knowledge base. Indeed the CSDSs contain information about services and their business and technical characteristics gathered and proven by previous experience. The requirements analysts do not need to start their analyses from the very beginning but to utilize the characteristics identified during the previous analyses. If the analysts identify some extra requirements, they could inquire about them the providers of the services they are interested in, through the CSDSs. We consider that in such a way new characteristics could be dynamically added depending on the current needs. Information maintenance would be a joint responsibility of both the service providers and service consumers.

In the service discovery phase the Cloud services that fulfil the identified requirements are looked up through CSDSs. We reckon that having a CSDS available, it is not necessary service discovery to start only after all requirements have been clarified. Instead of that both phases could run iteratively in parallel. As a result this will increase the process efficiency.

During the third phase service provider and service consumer negotiate a service level agreement. It is hardly believable this to be supported by the CSDSs since such kind of information is business critical for service providers and they will not be likely to share it publicly.

In the service composition phase several services provided by one or more providers are combined in a new service. The CSDS could be significantly helpful in automating this process as it contains all the necessary information about services' functionality and interfaces.

It is hardly achievable the last phase, that of service consumption and monitoring, to be part of the CSDS as the monitoring information is too sensitive to be publicly available. More probably, the consumers will receive such information directly from their providers. On the other hand, if the CSDS contains monitoring information, that will allow service quality to be objectively measured and, as a result, will improve the quality of the searching process. Additionally, it will be useful if consumers and providers give feedback about their satisfaction with the services.

Another CSDS's usage pattern is when enterprises or individuals intend to start using Cloud services and subscribe for receiving information about Cloud services of potential interest. When the CSDS identifies such services it notifies the subscriber who will analyse them.

Although the consumers may have already chosen appropriate Cloud services, it is likely that they will want to be informed if better services, in terms of functionality, performance or cost, appear. In such cases the consumers subscribe in the CSDS for being notified when this happens.

If the consumers are not satisfied with the quality of the running services they may want to change the providers. In that case the consumers consult the CSDS in order to find an alternative solution.

3.3 Functional Requirements

Functional requirements define the functional capabilities which CSDSs provide to the users – service providers and service consumers. We analyse the functional requirements in three aspects: (1) functionality of the system; (2) information about Cloud services, i. e. what kind of information about Cloud services is provided by the system; and (3) information provisioning, i. e. ways of communicating with the system. A detailed schema regarding these aspects is shown in fig. 3.

We consider that it is important CSDSs to provide information about both functional and non-functional characteristics of Cloud services. The existence of a complete and accurate list of services' features ensures that service consumers will be able to provide services to their customers with the required quality.

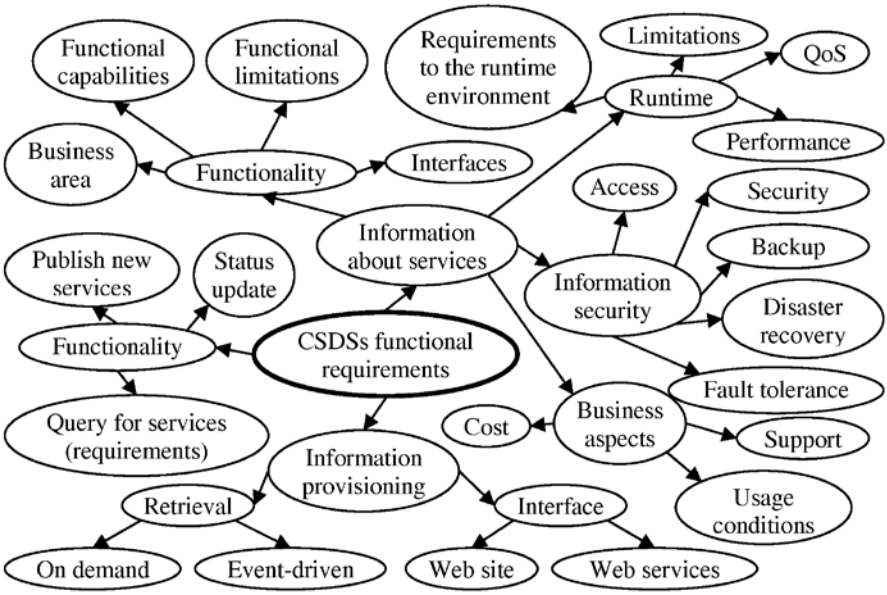


Fig. 3. Requirements for Cloud service discovery systems.

The publishing of information about some non-functional characteristics, especially regarding performance and security, may be a challenging task for

some service providers as they may be vulnerable to competition. However with the increasing of the quality of the offered services, providers will have stimulus to make such information publicly available as they could gain a competitive advantage.

3.4 Non-functional Requirements

Non-functional requirements, as usability, availability, reliability, etc., determine the CSDSs' quality. Some of the primary quality constraints for CSDSs are: (1) usability, i. e. the ease of use, on the one hand, and the consumers' satisfaction, on the other hand; (2) availability; (3) reliability, i. e. resistance to system failures; (4) scalability, i.e. the system to remain usable with growing number of service discovery requests; (5) response time; (6) performance; (7) security; (8) backup and disaster recovery; (9) extensibility, i. e. the ability to extend the system with new functionality; and (10) maintainability.

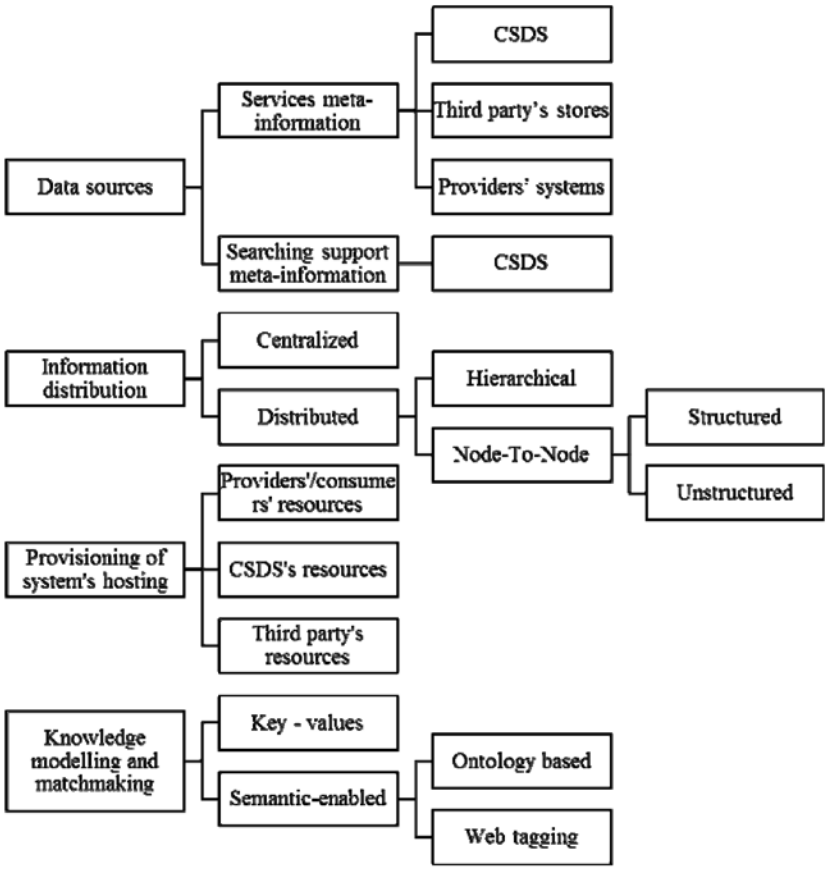


Fig. 4. Taxonomy of Cloud service discovery systems.

4 Architectural Models

In order to analyse the CSDS's architectural models we designed a taxonomy of CSDSs regarding their most essential architectural features. A schema of the taxonomy is present in fig. 4. The taxonomy's characteristics are described in details in the following subsections.

4.1 Data Sources

The data sources feature defines where the CSDSs get the necessary meta-information from. We regard this meta-information as composed of two kinds of information: (1) meta-information describing Cloud services themselves, such as functionality, performance, requirements, interfaces, etc.; and (2) meta-information that supports the searching process. In fig. 5 we present a layered model of the CSDSs regarding data sources.

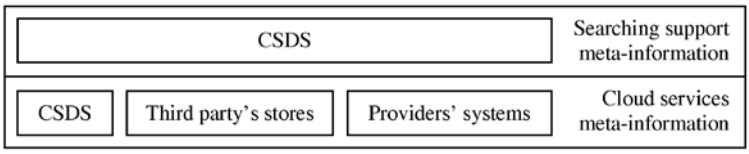


Fig. 5. Layered model of data sources for Cloud service discovery systems.

The first layer determines the meta-information which describes the Cloud services themselves. We identify a few possible data sources of that: (1) the meta-information is stored in the CSDS itself; (2) third parties' stores are used for initial information retrieval, e. g. conventional web search engines as Google, Yahoo, etc.; and (3) the meta-information is stored in providers' systems. The second layer adds to the systems various searching techniques and supporting meta-information. The features of the CSDSs implementing the abovementioned kind of data sources are given in the following paragraphs.

CSDS. All the meta-information is stored in the CSDS itself. Such systems allow information to be structured in a specific way in order to ensure more efficient searching. Service providers have to register services and their characteristics in the CSDS and to maintain the information up to date. Although this is the most accurate and efficient approach, in fact it is hardly achievable as it requires additional efforts by service providers. However the service providers will be encouraged to do that if (1) registering and maintenance are simplified, and (2) services become more attractive to the consumers because of availability of detailed information.

A typical CSDS of this kind is shown in fig. 6. The presented system is a

distributed one, which is the most common case for such kind of systems. In fact the schemas for centralized systems are very similar to the presented one except that they are hosted only on a single node. As a pre-condition for this schema it is supposed that service providers have registered their services in the CSDS. The data is stored on the respective node/nodes in accordance with the system's architecture.

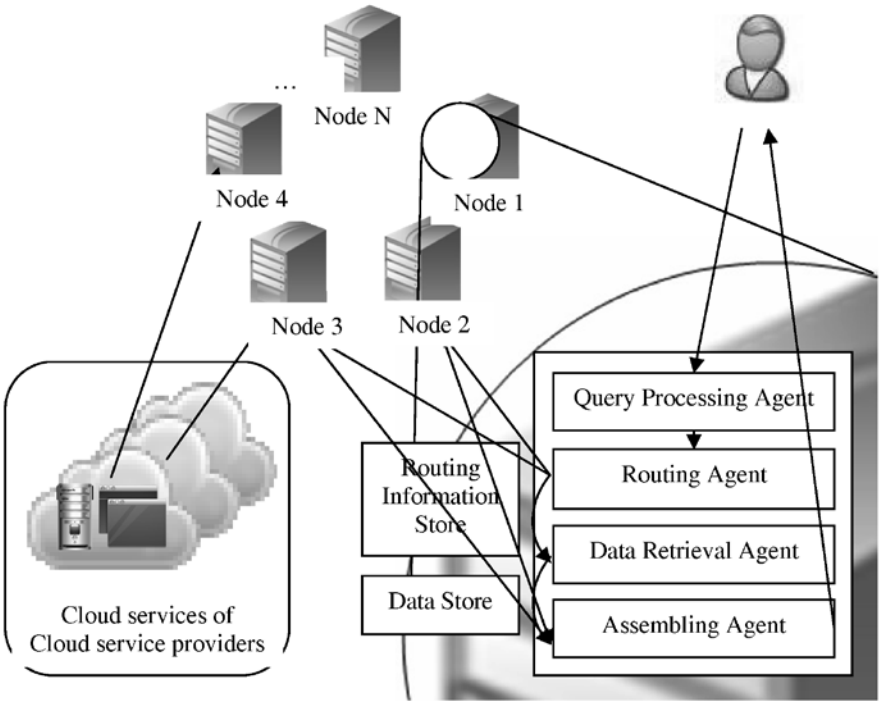


Fig. 6. Searching process in the CSDS which contains all the necessary meta-information.

The process of searching in principle is as follows: (1) the user sends a query to the CSDS to find a Cloud service by specifying the requirements to this service. The requirements could be specified either in a structured form, if the CSDS supports structured description of Cloud services, or as a list of key words. The query is transmitted for execution to the nearest node (or someone else); (2) the Query Processing Agent (QPA) on that node validates the query and transmits it to the Routing Agent (RA); (3) the RA consults the local Routing Information Store in order to determine which system's nodes contain the searched information. The RA divides the query into subqueries – one for each corresponding node; (4) the RA sends the subqueries to the nodes; (5) the Data Retrieval Agents on these nodes retrieve the requested data from the local data stores and return the results

back to the requesting node; (6) the Assembling Agent of the requesting node joins the received information and returns the result to the user.

Third parties' stores. The CSDS does not store information about Cloud services themselves. Instead of that third parties' stores, for example conventional web search engines as Google, Yahoo, etc., are used for initial information retrieval. The returned data is then filtered in order to find the most reasonable results using Cloud ontologies or other heuristics. This approach cannot ensure high quality of the found information but it is easy for maintenance by the service providers as the information is taken from the Internet.

A typical CSDS of this kind is shown in fig. 7. Such systems are rather centralized. As a pre condition for this schema it is supposed that service providers have published information about provided Cloud services in the Internet (in corporate web-site or any other web-sites).

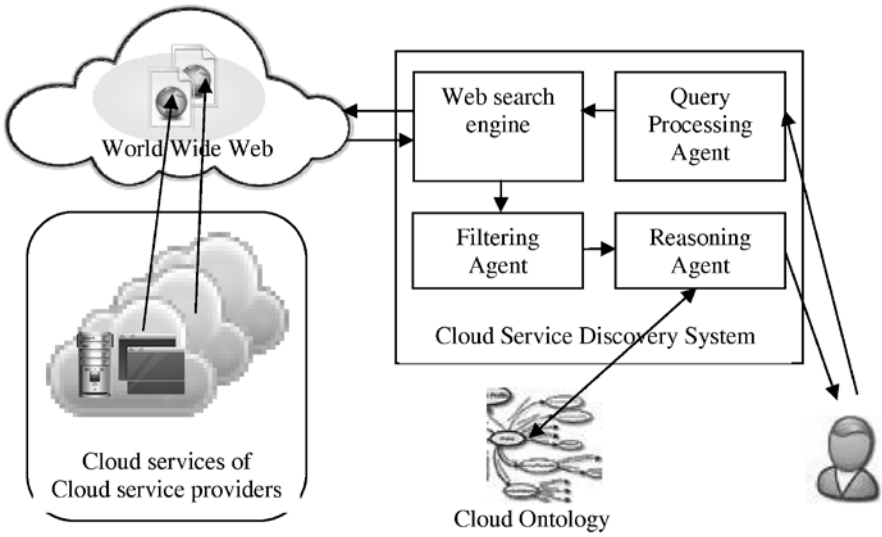


Fig. 7. Searching process in ontology-enhanced web search engines.

A typical searching process consists of a few steps: (1) the user sends a query to CSDS for searching of a Cloud service that fulfils certain requirements. These requirements could be specified either in a structured form, following the Cloud ontology's concepts, or as a list of key words; (2) the Query Processing Agent (QPA) processes the query and composes a search expression consisting of the key words from the requirements; (3) the QPA submits the search expression to a web search engine and transmits the result set to the Filtering Agent (FA). If

the result set contains very few web-pages, a new alternative query is generated in order to return more results; (4) the FA filters the most relevant web-pages using various heuristics, e. g. counting the frequencies of evidence phrases in the web-page content; (5) the Reasoning Agent (RA) checks the semantic similarity between the request and the found web-pages. In order to do that the RA at first transfers the query into concepts of the Cloud ontology; then discovers the concepts presented in each of the web-pages; and finally calculates the similarity between both types of concepts using various similarity reasoning techniques. The RA ranks the web-pages according to the calculated ratings; and (6) the CSDS returns to the user the web-pages with the highest service utility.

Providers' systems. Each service provider develops and supports its own local information service which gives information about provided Cloud services, their characteristics and state upon request of the CSDS. The CSDS serves as an additional layer which transparently maps the individual information services into a single federated system. A real obstacle to such systems is that in general they have to support multiple interfaces with service providers, which makes information unification difficult.

The searching process is similar to the one demonstrated in fig. 6 except that the Routing Agent sends the subqueries to the respective providers' information services.

4.2 Information Distribution

This aspect characterizes whether information is hosted on a centralized site or is distributed across many sites.

Centralized. Centralized systems have good performance and simple maintenance but they suffer from some drawbacks, e. g. single point of failure, lack of scalability, strong network bandwidth to the node which hosts the system, etc.

Distributed. Distributed systems are highly scalable, can gracefully adapt to the dynamic system expansion or contraction and outages, and avoid single point of failure.

Distributed/Hierarchical. Hierarchical systems are organized into a tree-like structure. Although they are appropriate for searching by a single attribute they are not naturally designed to support multi-dimensional queries. Another disadvantage of such systems is that in case of data update the hierarchy has to be reorganized which could cause serious performance problems with highly dynamic data.

Distributed/Node-To-Node/Structured. Structured systems use a rigid structure

to organize an efficient search overlay network. The basis of this structure are distributed indices usually based on Distributed Hash Tables (DHT). DHTs offer deterministic query search results within logarithmic network message complexity. They have good efficiency for single-dimensional exact queries such as "finding all services which have a specific value for a predefined attribute" but supporting multi-dimensional range queries, which in fact is the most common case in CSDS, is a more complex and challenging task.

Distributed/Node-To-Node/Unstructured. Unstructured systems are characterized by the lack of an underlying organizational structure. There is no deterministic information about resource location in the system. Therefore the prevailing request propagation method is controlled flooding – a process in which requests are forwarded until their Time-to-Live parameter is exhausted. Unstructured systems offer great flexibility in adding new characteristics and are simple for maintenance.

4.3 Provisioning of system's hosting

There are few business models for that.

Providers'/consumers' computational resources. Service providers or service consumers or both provide computational resources on which the CSDS is hosted. Providers have reason to support technically and financially CSDSs as they provide them with efficient channels for supplying Cloud services to consumers. Consumers' arguments are that CSDSs could save a lot of efforts and time in searching appropriate Cloud services and ensure the best possible selection. The CSDSs' usage could be free of charge as the significant part of the operational expenses, those for the computational resources, is covered by providers and consumers themselves.

CSDS's or third parties' computational resources. Hosting computational resources are provided by the service broker or third party. They could be provided as conventional computational resources or as Cloud services (IaaS). Service providers or service consumers or both could pay some fees to use the system or the usage could be free of charge and the operational expenses could be covered by advertising.

4.4 Knowledge Modelling and Matchmaking

This may include semantic annotations and web tagging for example through suitable keywords, or the simplest case, attributes in the form of key-value pairs. Another approach is to use ontologies in order to store and correlate the information from various sources. This helps not only in determining the important keywords

but also in classifying hierarchically the numerous concepts that may be used to describe a service or resource.

5 Conclusion and Future Work

Availability of efficient service discovery techniques is a crucial factor to the success of Cloud computing. CSDSs help service consumers in their decisions about usages of Cloud services or changing service providers. Defining business and architectural constraints is the first step in designing efficient CSDSs. In order to do that we proposed a systematic approach based on a multi-criteria decision making. Following this approach our contributions are: (1) we analysed and determined the usage patterns and processes in which CSDSs are involved; (2) we defined the most essential functional and non-functional requirements for CSDSs; and (3) we designed a taxonomy in order to determine an appropriate architectural model.

Our future work will be focused on designing a CSDS, based on the identified requirements, which achieve reasonable balance between usability and efficiency, simplicity and reliability.

References

1. Bhama, S., Karpagam, G. R.: Realizing the Need for Similarity Based Reasoning of Cloud Service Discovery. International Journal of Engineering Science and Technology, vol. 3, pp. 8404-8414 (2011)
2. Fernandez, A., Hayes, C., Loutas, N., Peristeras, V., Polleres, A., Tarabanis, K.: Closing the Service Discovery Gap by Collaborative Tagging and Clustering Techniques. Proceedings of the Second Internatioal Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web (SMR2 2008), Karlsruhe, Germany, (2008)
3. Gartner Press Release "Gartner Says Worldwide Software-as-a-Service Revenue to Reach $14.5 Billion in 2012", http://www.gartner.com, March 27 (2012)
4. Han, T., Sim, K. M.: An Ontology-enhanced Cloud Service Discovery System. International MultiConference of Engineers and Computer Scientists (IMEC 2010), Hong Kong, pp. 644-649 (2010)
5. Joshi, K., Finin, T., Yesha, Y.: Integrated Lifecycle of IT Services in a Cloud Environment. Proceedings of the Third International Conference on the Virtual Computing Initiative (ICVCI 2009), Research Triangle Park, NC (2009)
6. Kang, J., Sim, K. M.: A Cloud Portal with a Cloud Service Search Engine. International Conference on Information and Intelligent Computing 2011, IPCSIT vol.18, IACSIT Press, Singapore (2011)
7. Kousiouris, G., Kyriazis, D., Varvarigou, T., Oliveros, E., Mandic, P.: Taxonomy and State of the Art of Service Discovery Mechanisms and Their Relation to the Cloud Computing Stack. Achieving Real-Time in Distributed Computing, chapter 5, pp. 75-93, IGI Global (2011)
8. Ranjan, R.,Zhao, L.,Wu, X., Liu, A.: Peer-to-Peer Cloud Provisioning: Service Discovery and Load-Balancing. Cloud Computing, Computer Communications and Networks vol. 0, pp. 195-217, Springer London (2010)

9. Zhelev, R., Georgiev, V. Resource Information Service for Cloud Datacenters. Proceedings of the International Conference on Information & Communication Systems, Irbid –Jordan, May 22-24., pp. 83 – 88 (2011)
10. Zhelev, R., Georgiev, V. A DHT-based Scalable and Fault-tolerant Cloud Information Service. Proceedings of the UBICOMM 2011: The Fifth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Lisbon – Portugal, November 20-25, pp. 66 – 72 (2011)
11. Zhou, J., Abdullah, N. A., Shi, Z.: A Hybrid P2P Approach to Service Discovery in the Cloud. International Journal of Information Technology and Computer Science (IJITCS), vol. 3, no. 1, pp.1-9 (2011)

Monitoring of Business Processes in the EGI

Radoslava Hristova

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

radoslava@fmi.uni-sofia.bg

Abstract. The European Grid Infrastructure (EGI) uses partially service-oriented grid middleware for grid computing (g-Lite). In the context of SOA-based business process management the definition, monitoring and optimization of a business process in the infrastructure are still not supported. In this article we present an approach for business process monitoring in EGI, based on the service-oriented BPM platform.

Keywords: grid, monitoring, business processes

Introduction

The business process is a "set of logically-related tasks performed to achieve a defined business outcome" [1]. Business prosess management (BPM), "supports business processes using methods, techniques, and software to design, enact, control, and analyze operational processes involving humans, organizations, applications, documents and other sources of information" [2].

The service-oriented architecture (SOA) [3] is an architectural style for developing systems and applications. Basic characteristics of the model are well-defined logical entities called services, which can be independently used. Applaying service-oriented style for BPM will improve the design and management of the business processes and will optimize their usage. Combining business process management with the service-orientated architecture will provide flexibility and optimization to the developed business processes and reuse of existing assets. Evenmore the service-oriented approach is the preffered approach for building cloud systems – the next grid generation. In [4] the authors proposed such service-oriented generic resource framework for cloud systems, which represents datacenter resources in a uniform way, allowing generic administration without knowledge of the underlying resource access protocol.

The SOA lifecycle (Figure 1) consists of four phases: model, assemble, deploy and manage. We are describing them with respect to business process monitoring. During modeling phase are gathered requirements. In this phase the business processes are designed. If the business process will be monitored, key

performance indicators (KPIs) are defined. Evenmore, simulations of the business process with the defined key indicators can be done. During assembe phase, new assets are developed or existing assets are used. The business processes are composed and the key performance indicators are implemted. During the deploy phase the developed business processes and key indicators are deployed into the process server and monitoring server. Thus, the people, process and information are integrated. In the last phase of the cycle the developed business processes, services and applications are managed. The defined business metrics are monitored. If during the SOA lifecycle, some improvements can be done, the business process is optimized and processed through the lifecycle again.

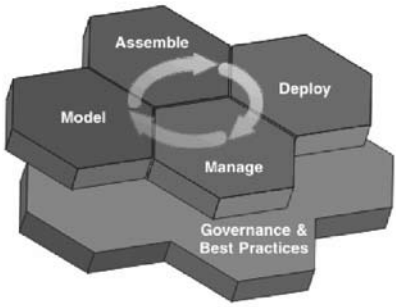


Fig. 1. Service-oriented architecture lifecycle

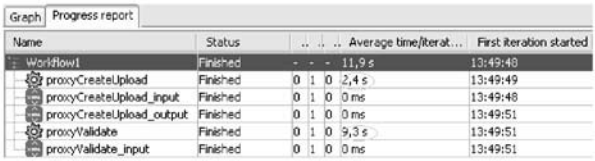
The European Grid Infrastructure [5] (EGI) uses partially service-oriented grid middleware for grid computing (g-Lite) [6]. In the context of SOA-based business process management the definition, monitoring and optimization of a business process in the infrastructure are still not supported. In [7] the author discusses the important aspects of service-orientated grids and underlines the lack of widely accepted mechanisms for business process orchestration, mediation and monitoring in it. G-Lite is not an exception. The goal of current investigation is to present an approach of business process monitoring for the EGI, based on service-oriented BPM platform for the EGI.

Tools in the EGI and their service-orientation

In [8] we present some of the tools, which are available for g-Lite and can be used for building and executing service compositions in the EGI. All of the presented tools use their one mechanism for service composition, independently from g-Lite. We will focus on three of these tools, which fulfil the requirements for service-orientation: Triana [9], Taverna [10] and Kepler [11]. Our choise is influenced and from [12], where the same tools are described as grid tools for monitoring and control of the workflow execution. We are describing the tools with respect to the features they provide for business process monitoring.

The software Triana is a distributed environment, which provides functionality for building and executing business processes (workflows). The software is adapted for access to the grid resources, including the g-Lite middleware. From architectural point of view, the environment consists of three major components: user interface, Triana service and Triana control service. According [13] Triana could be used as a visual environment for monitoring the workflow of grid services. The user can get the information about the time for which the workflow is executed and the information about the status of the job (done, running, etc.).

Taverna software provides environment for business process design and execution in grid. From architectural point of view, the software consists of two major modules: Taverna Workbench and Taverna Engine. The Taverna Workbench is a graphical editor for business process modeling, which provides functionality for monitoring of the designed process. On (Figure 2) is shown example of progress report for workflow executed into the EGI. The monitoring gives information for the status and the average time for execution of the workflow.



Graph | Progress report

Name	Status	..	..	..	Average time/iterat...	First iteration started
Workflow1	Finished	-	-	-	11,9 s	13:49:48
proxyCreateUpload	Finished	0	1	0	2,4 s	13:49:49
proxyCreateUpload_input	Finished	0	1	0	0 ms	13:49:48
proxyCreateUpload_output	Finished	0	1	0	0 ms	13:49:51
proxyValidate	Finished	0	1	0	9,3 s	13:49:51
proxyValidate_input	Finished	0	1	0	0 ms	13:49:51

Fig. 2. Taverna monitoring tool

Kepler is a tool for design of business processes (workflows). The tool provides a graphical interface for business process modeling. The business processes can be described by using five basic components: directors, actors, parameters, link and ports. The director controls the execution of the process. The actor provides realized functionalies and follows the director's instructions. Every process has exactly one director. In [14] authors discussed how grid workflow designed in Kepler can be monitored. The monitoring process gives information for the jobs that are executed on the grid. The users should be able to view the status of the workflow (submitted, active, done, etc.), the tasks currently executing, and other information.

The three envirionments Triana, Taverna and Kepler support service-orientation and can be used for workflow execution in the EGI. Unfortunattly, the monitoring tools wich these envirionments provide, do not allow user to customize, define or choose his own indicators for monitoring. That is not the case for SOA-based BPM systems. In the next section we present such system and example solution for this problem.

SOA-based BPM Platform for EGI

In [15] we present a framework for service-composition in the EGI. The software implementation of the framework is based on a SOA-based BPM Platform, presented on (Figure 3).

On the first layer of the model are the legacy EGI services and applications. The layer covers already built it grid infrastructure, i.e., all the EGI services and applications are available and they can not be changed.

On the second layer are all of the developed web services for access to the EGI services and application. The web services are important part form the framework. They participate into compositions in the higher layers. Additional layer is implemented in order to do the model applayable for the EGI. The module provides adapters for access to the EGI services and applications and exposed them as web services. The current implementation of the module provides web services for access to g-Lite and web services for access to ROOT application. The developed web services are composable.

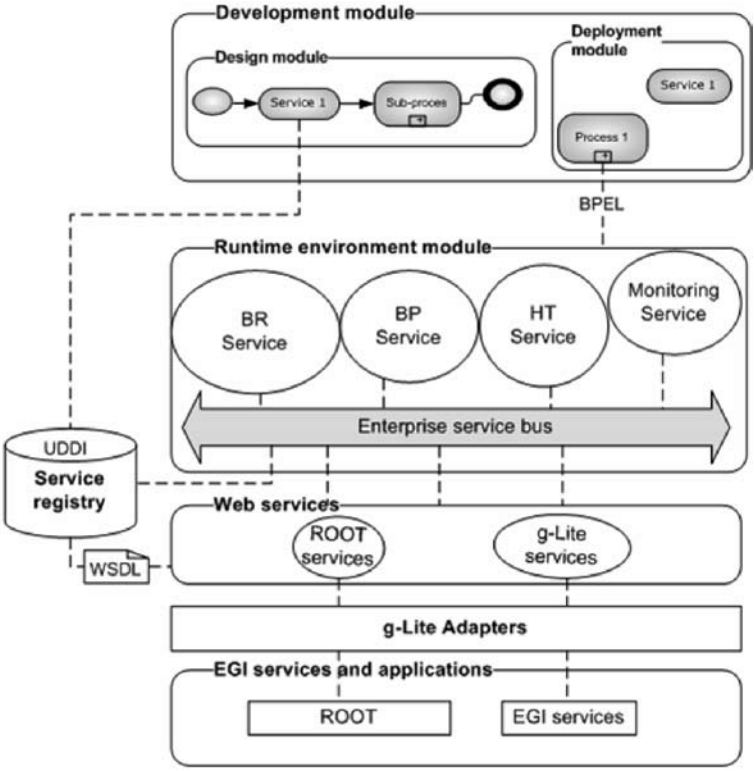


Fig. 3. Architecture of a SOA-based BPM platform for the EGI

On the third layer are the registry services, which provide features for publishing and discovery of the developed web services from the second layer. The access to the service registry can be done through the development module and through the runtime module. In [16] the authors present comparison of grid resource discovery approaches. Based on the functional requirements that they defined, we can conclude that, the UDDI approach is the most appropriate approach for grid resource discovery with respect to the defined criterions. The SOA-based BPM system also relies on the UDDI. Our implementation supports the UDDI solution for service registry.

On the forth layer is the runtime envirionment module of the framework, which includes the enterprise service bus, business-process management service, business process monitoring services, human tasks services and business rules management services.

The fifth layer of the framework includes development module. It covers instrumental tools for design and deployment of the business process into the runtime envirionment module. The business processes can be designed, executed, managed and monitored in the forth and the fifth layers from the framework.

In this article we are focussing on the monitoring services from the runtime envirionment module of the framework. For the implementation of the service we used IBM WebSphere Business Monitor tool [17].

IBM WebSphere Business Monitor is business activity monitoring software that provides monitoring of events in real time by providing visual display of business process status, together with alerts and notifications. The users can create KPIs without requiring a new IT development and deployment cycle. They can expand KPI calculations, including calculations based on the relationship with other KPIs and also can receive automated alerts for warnings.

The monitoring of the business process starts from the design module, where the business process is designed. The grid user can set KPIs during modeling phase. After that the business process is translated to the assemble phase, where some business assets are reused or implemented. For the defined KPIs in the modeling phase KPIs in the assembly phase are generated. Examples of KPIs in the deployment module are shown on (Figure 4). The grid user can choose either to monitor the start time of the process, the average working duration or the end time. With this solution more monitoring information can be collected and more information can be recieved.

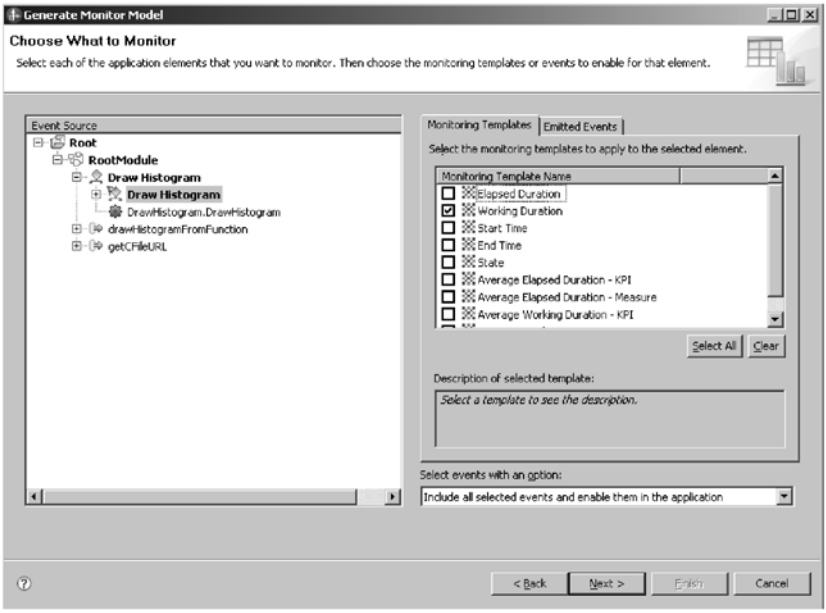


Fig. 4. Monitoring templeates for KPIs for the business processes

The developed business process is deployed into the runtime envirionment, where the monitoring service and business process management service executes the deployed business process. The result from the excution and monitoring is displayed into the business space portal which is part form standart distribution of the IBM WebSphere Business Monitor tool. Example output is shown oh (Figure 5)

Fig. 5. Business space – business process monitoring

The gauges are one of the ways to visualize information by representing KPI values on a gauge. A dial is used to represent the position of the KPI value relative to the range and target of the KPI. A needle indicates the current value of the KPI. The gauge view focuses on representing KPIs that belong to aggregated business measures in a monitor model. Each gauge represents the value of a single KPI.

Conclusions

In this article we present an approach for business process monitoring in the EGI, based on the service-oriented BPM platform. With this solution more monitoring information can be collected and the grid user can receive more information, which can be used for improvements and optimisations of the business processes.

Acknowledgement

This paper is supported by Sofia University "St. Kliment Ohridski" SRF under Contract 134/2012.

References

1. Davenport, T., Short, J. "The New Industrial Engineering: Information Technology and Business Process Redesign", 1990
2. van der Aalst, M., at all., Weske, M. "Business Process Management: A Survey", 2003, http://bpt.hpi.uni-potsdam.de/pub/Public/PaperArchive/bpm2003.pdf
3. Keen M., at all., Patterns: SOA Foundation - Business Process Management Scenario, International Technical Support Organization, 2006, http://www.redbooks.ibm.com/redbooks/pdfs/sg247234.pdf
4. Zhelev R. and V. Georgiev. A Generic Resource Framework for Cloud Systems, Proceedings of The 4th International Conference on Distributed Computing and Grid-technologies in Science and Education, June 28 - July 3, 2010 Dubna, Russia., pp. 268 – 278.
5. European Grid Infrastructure, http://www.egi.eu/
6. g-Lite, http://glite.cern.ch/
7. Dimitrov, V. T., Development of applications with service-oriented architecture for grid, ACM New York, 2008, Proceedings of the 9th International Conference on Computer Systems and Technologies (CompSysTech '08), Article No.14
8. Goranova R. D., Service composition tools in g-Lite, Conference Proceedings of the Fifth International Conference ISGT, 2011, pp. 228-235
9. Triana 4 User Manual, http://www.trianacode.org/docs/userguide/UserGuide.pdf
10. Taverna 2 Architecture, http://www.taverna.org.uk/developers/taverna-2-x/architecture/
11. Getting Started with Kepler, https://code.kepler-project.org/code/kepler-docs/trunk/outreach/documentation/shipping/2.3/getting-started-guide.pdf
12. L. Kirchev, V. Georgiev and K. Boyanov, Workflow Management for a General Purpose Grid Platform of Commodity Computers, in Proceedings of the International Workshop on Network and GRID Infrastructures, Sofia, Bulgaria, 27-28 September, 2007. pp. 42 – 50.
13. Wang, I., Taylor, I., at all. Triana as a Graphical Web Services Composition Toolkit, Cardiff University, 2003
14. Nandita Mandal, N., Deelman, E., at all. Integrating Existing Scientific Workflow Systems: The Kepler/Pegasus Example, USC Information Sciences Institute, 2007, http://pegasus.isi.edu/publications/kepler-works07.pdf

15. Goranova, R. D., Framework for service composition in g-Lite, American Institute of Physics, Conference Proceedings Volume 1404, 2011, pp. 218-224, ISBN 978-0-7354-0976-7, ISSN 0094-243X.
16. Pashov G., K. Kaloyanova, Comparison of Grid Resource Discovery Approaches, Third International Conference on Information Systems & Grid Technologies, 28 - 29 May 2009, Sofia, Bulgaria, pp 138-147
17. WebSphere Business Monitor V6.0.2 - Features and usage scenarios, IBM Corporation, 2007

Computational Challenges in a Metagenomic Data Processing Pipeline

Milko Krachunov[1], Dimitar Vassilev[2], Ivan Popov[3], Peter Petrov[1]

1 Faculty of Mathematics and Informatics, Sofia University "St.Kliment Ohridski" 5 James Bourchier Blvd., 1164 Sofia,.Bulagria,
2 Bioinformatics group, AgroBioInstitute, 8 Dragan Tsankov Blvd., 1164 Sofia, Bulgaria,
3 Molecular Medicine Center, Medical University-Sofia, 2 Zdrave Str., 1431 Sofia, Bulgaria.

Abstract. Researchers in metagenomic studies are faced with a variety of difficulties, including the shortage of tools, computationally expensive data processing, uncertain dataset quality, and the need to do extensive empirical data analysis. Our project, which was focused on the development of an efficient error detection approach, revealed a need to offer a solution to some of the other problems as well.

Keywords: metagenomics, highthroughput sequencing, error detection algorithms

1 Introduction to metagenomics sequencing studies

The essence of any living organism is described in its DNA and RNA molecules which are the primary carriers of the genetic instructions that control its development and all the biological processes that occur throughout its life. As the information they contain is chiefly digital in nature, from the point of view of an informatician a DNA molecule is nothing but a very long sequence of a four-letter alphabet. With the rapid advancement of DNA sequencing technologies, the amount of such sequences available to us for analysis is increasing exponentially.

Once acquired, they provide a multitude of opportunities for research. Some studies focus on finding and studying the expressive, meaningful of those sequences - the genes producing results significant over multiple branches of sciences. Other studies produce useful results relying only on superficial analysis of the sequence structure - variations alone can be used to identify individuals in DNA profiling performed by forensics [1], or to discover the evolutionary relationships between species in phylogenetics. [2]

An example of the latter kind is metagenomics. It is a new and largely unexplored field in genetic research that deals with mixed genetic material found in environments ranging from soil and water basins to the insides of various

macro-organisms. A single metagenomic dataset contains the sequences of a large number of organisms, mostly microbes, from a variety of species. Commonly pursued tasks can be the comparative analyses of microbial communities for the interest of human health [4] and agriculture, or the studies of the bacterial and viral evolution. [3] As the the most rapidly mutating agents in our biosphere, microbes can provide a lot of insight on evolution, and are also a critical factor in unexpected disease outbreaks.

A researcher in the field of metagenomics is confronted by a variety of challenges. On one hand, as a new field there are yet no well-established methods to approach it, and it is not uncommon to face an unsolved technical or methodological problem. On the other hand, the nature of the data itself does yield well to analysis. Not only the datasets are big and heterogeneous, but most of the microbial species comprising them are unknown, and because of their vast number and their rate of evolution, cataloguing them all might not be a feasible task. This leads to deficiencies in both the quality of the data, and the possibility for computational optimizations.

As a result, a great deal of the work involves computationally expensive and rigorous processing of huge datasets and a great deal of uncertainty about the correctness of the data that is being processed.

2. A case study

To illustrate the inherent difficulties in metagenomics, we will present our work on a few datasets. The goals of our project are:

- To produce a software package that eases the execution of metagenomic experiments.
- To develop and test an approach for error detection and correction integrated in that software.
- To fill the gaps in the available software packages that we stumble upon.

2.1. The input data

A common sequencing target for bacteria is the 16S RNA. It is very attractive for metagenomic data analysis, because it is highly conserved and thus largely preserved across a great deal of species, while at the same time it contains a hypervariable region that is incredibly helpful for identifying species, individual organisms and tracing their evolutionary relationships. [2]

Our sample datasets contain 16S RNA short sequences that are 300 to 500 bases in length, and each set comprises tens of thousands (20000 to 50000) of reads after filtering.

2.2. Sequence alignment

Sequence alignment is a crucial step before performing any further data analysis. Let us look at a sample excerpt from one of our datasets:

TCTCTATGCGCCATTGTAGCACGTGTGTAGCC …
TCTCTATGCGCCATGTAGCACGTGTGTAGCC…

One can easily notice that the second sequence is displaced because a base at position 15 is missing. It is impossible to perform any meaningful column-wise analysis unless such displacements are accounted for. Certain experiments can avoid the expensive sequence alignment by using less reliable comparisons based on k-mers, but in most cases this is not a viable option, and gaps need to be added to the shorter sequences.

Unfortunately, finding the globally optimal alignment for n sequences is an NP-complete problem. For any considerably sized dataset, finding this optimum is a practical impossibility.

For other genomic studies, this problem is not very pronounced, because the sequences are usually limited a single species and the test sequences are most often aligned one by one against an already known representative sequence for that species. With metagenomics,neither of these is an option, and the datasets are considerably bigger. This means that the vast majority of software designed for sequence alignment is unusable for metagenomics.

There are several practical approaches that produce a reasonable alignment in a reasonable amount of time. One of the most commonly used is based on guide trees and has a time complexity of O(n3).

Algorithm 1: Guide-tree sequence multiple sequence alignment

1. *A distance matrix of the sequences is produced by aligning each pair of sequences along each other.*
2. *Using a neighbour-joining hierarchical clustering, the sequences are placed in a binary guide tree. This tree is similar to a evolutionary tree, and indeed, the algorithm used to construct it is the same, but the latter is distinct because it uses a different type of distance matrix. This is the O(n3) portion of the algorithm.*
3. *Starting from the leaves of the tree, each pair of sequences is aligned against each other, and the node between them is replaced with a sequence pattern representing both of them, then these patterns are aligned up until the root.*
4. *The gaps that were introduced in each path from the root to a leaf are the gaps that are added to the original sequences to produce the final alignment.*

In our experiments we discovered that alignment software packages that were meant for large datasets were either unable to process our data in a reasonable time, or unable to produce acceptable results. It should also be noted that a direct implementation of algorithm 1 did not lead to better results, because the solution it gives is often very far from the optimal one and requires further computationally expensive improvements to be applied.

To remedy this we used a very simple and straight-forward approach. We did a quick rough clustering of the dataset into a dozen of clusters using CD-HIT [8,9]. We aligned each cluster with a software solution producing a high-quality alignment in reasonable time, in particular MAFFT [5] and MUSCLE. [6] Then, we aligned the clusters against each other in a manner similar to steps 3 and 4 of algorithm 1.

Surprisingly, the alignment took significantly less time and was significantly superior in quality to the alignment produced directly by MAFFT and MUSCLE with low-quality settings.

While evidently the alignment of the metagenomic datasets is feasible, we could not find a straightforward solution and were forced to improvise, even though alignment is a very basic component of the metagenomic processing.

2.3 Error detection and correction

2.3.1 Problem description

Another obstacle in many metagenomic studies is the uncertainty about the data correctness. There are two kinds of errors that can occur in your metagenomic dataset - errors produced by the sequencing equipment, and errors from biological origin, i.e. mutations.

While initially both occur randomly, the mutations most often lead to an evolutionary dead-end killing the organism, which makes them uncommon in your sample. Any mutations that do remain can carry some particular information about the species or the individual that makes an important study target. At the same time, the two are very difficult to tell apart without specific knowledge about their function.

The only way to distinguish errors is by their frequency of appearance, but that is not entirely reliable. It is a common practice to throw out any reads for which there is a suspicion of errors, which can reduce the dataset by an order of magnitude, and in most this leaves the suspicion that some of the discarded data was actually correct.

As errors are the second most critical problem of metagenomic studies, one that is largely unsolved and possibly unsolvable, any progress with it would

greatly affect any metagen omics study, and error detection is an important part of any software distribution aimed at metagenomics.

2.3.2 The naпve approach

The most obvious way to find errors is simply look for bases that occur rarely. You need to count the frequency of occurrence of each base at each column, and the bases that appear less frequently than a previously established threshold can be considered an error. This is based on the assumption that while mutations are rarer, a single mutation would appear more often than an error because it has survived through multiple generations. Another expectation is that most organisms will appear multiple times in the sample by chance.

To correct the errors, one might naпvely replace the offending base with the one that appears most often in this column. And, indeed, this approach makes a lot of sense if the dataset is for a newly-sequenced genome of a single organism. It is entirely equivalent to the established sequencing approach - you acquire enough reads to produce a reasonable coverage of the region in question, and for each position, select the base that is occurring most often. [7, 10]

Unfortunately, this makes no sense when the dataset contains multiple distinct organisms.

For each position, you might have multiple competing options that are all correct. And you might have entire reads that occur less often than the threshold that are also correct.

If R is the set of reads of size n, the naive approach score for position k in read r can be expressed mathematically as:

$$\text{score}(r,k) = \sum_{p \in R}^{p \neq r} \frac{[r_k = p_k]}{n-1} = \frac{\sum_{p \in R}^{p \neq r} [r_k = p_k]}{\sum_{p \in R}^{p \neq r} 1} \quad (1)$$

2.3.3 Similarity based approach

To make the error detection and correction more suitable for the heterogeneous nature of the dataset, we propose several improvements over the most basic approach. We would still count the frequencies of occurrence, but we will do so taking the context into account. We have the following requirements:

- A mismatch between two similar reads should be more important than a mismatch between two dissimilar reads. Or more generally, the importance of the mismatch should be proportional to the similarity between the reads.

- The similarity in the proximity of the mismatch should be more important than the similarity away from it. Or more generally, the importance of the similarity at a given position should be inversely proportional to the distance to the mismatch.

```
TCTCTATGCGCC ATTGT AGCACGTGTGTAGCC… (6716)
TCTCTATGCGCC ATAGT AGCACGTGTGTAGCC… (20)     <- p
TCTCTATGCGCC TCACG AGCACGTGTGTAGCC… (20)     <- r
TCTCTATGCGCC TCTCG AGCACGTGTGTAGCC… (1)
             i k
```

The reason we want to count only mismatches in similar reads is obvious - the probability that multiple errors occur alongside each other is significantly lower than a probability of a single error, which means that if the entire sequence or region is different, it is much more likely that it is an unrelated organism. There are two advantages to taking the similarity into account - if we find an entire region that has been replaced, we won't mistake it for an error, and if there are two distinct sets of sequences, we would separate them when we are trying to correct an error in one of them.

The proximity requirement should be also evident from that, but let's re-iterate that if two bases are close to each other they are much more likely to be functionally related to one another.

Algorithm 2: Similarity-based error detection

1. A window is created around each evaluated position.

2. For each two reads compared for a mismatch at that position, the similarity in the window is calculated. The positions right next to the evaluated one, have a higher weight in the similarity score.

3. The mismatch score for that pair and position is incremented with the similarity score divided by the sum of all similarity scores in the position.

If we amend the formula from (1) with our similarity score, we would get the following:

$$\text{score}(r,k) = \frac{\sum_{p \in R}^{p \neq r} \text{similarity}(r,p,k)[r_k = p_k]}{\sum_{p \in R}^{p \neq r} \text{similarity}(r,p,k)} \tag{2}$$

Choosing a good similarity function is tricky. Our proposal uses one that declines exponentially as we move away from the evaluated position. This has the effect that the window size becomes irrelevant after a certain point, as the importance quickly tends to zero. Let q be some parameter that has to

be experimentally evaluated and w be the size of the window that we deemed suitable, we can calculate the similarity as following:

$$\text{similarity}(r, p, k) = \frac{\sum_{i \in \text{window}(r,k)} q^{|m-i|}[r_i = p_i]}{\sum_{i \in \text{window}(r,k)} q^{|m-i|}} \quad (3)$$

where,

$$\text{window}(r, k) = \{i : \exists r_i\} \cap (\{k - w, \ldots, k - 1\} \cup \{k + 1, \ldots, k + w\}) \quad (4)$$

The time-complexity of our proposal is O(wn2), however, we can cache all the possibilities within a window. In theory, this leads to the useless time-complexity of O(neW), but if the window size is small and the actual variation per position is also small enough, it can be practically faster.

We should also mention a curious side-effect that the similarity-based approach has. Normally, when you are doing sequencing, you only use sequences from the particular organism to confirm a sequence. In a metagenomic dataset, however, different organisms are generally unavoidable. This is always undesirable, since it only leads to additional noise in the data. However, with a local similarity filter that noise will be largely filtered, and in rare occasions two different organisms might confirm each other for a region that was largely conserved between them.

For error correction, we can simply apply the same calculation for any potential replacement for the given position. This makes more sense than with the naпve approach, because multiple possibilities are much more likely to be grouped together.

2.4 Validating error detection and correction

One of the larger problems of metagenomic analysis is that it is impossible to obtain a reliable test dataset. Any test dataset you obtain will be either unrepresentative of a typical sequencing run, or will contain errors because of an error filtering applied. Even a set constructed from databases of known microbes will be biased when it comes to error detection, because it is likely that some of the mutated rare variants will be missing. And it is very difficult to obtain a set that resembles the distribution of microbes that you will get from an actual run by using data from a database.

So, while creating simulated reads using known microbe sequences is definitely a possibility, we will offer two more options for validation.

2.4.1 Repeated application of an evaluated error correction approach

While you can not reliably simulate reads, you can reliably measure and simulate

the errors that the sequencing equipment produces. In order to do so, you just need to run a set of genomic sequences that are known to you and then create an error profile. Once you have an error profile, you can apply the following procedure:

Procedure 1: Repeated application procedure

1. *Obtain dataset 0e from the sequencing equipment.*
2. *Apply the error correction approach to dataset 0e, producing dataset 1.*
3. *Simulate errors in dataset 1 according to the error profile that you had estimated, producing dataset 1e.*
4. *Apply the error correction approach to dataset Ie, producing dataset 2.*
5. *Measure the differences between datasets.*

At first it sounds counter-productive to introduce errors in a dataset that already contained errors and re-apply the error correction however this might produce several useful figures. If the error profile is correct, the comparison between 1, 1e and 2 gives an accurate estimate about the false negatives of the approach. The amount of missed simulated errors will be roughly the same as the amount of missed real errors. At the same time, the difference between dataset 0e and dataset 1 gives you the total amount of corrections made.

In a quick comparison of the naпve approach and the similarity-based approach in this manner, we measured a significant decrease in overall corrections made with no change in the number of the false negatives, which would suggest a decrease in the false positives.

A more extensive testing with this and the other approaches is still required to get more affirmative results.

2.4.2 Subset approach

The main difficulty in obtaining a test metagenomic dataset is the impossibility to sequence the exact same sample again. If that was possible, you would simply sequence the sample over and over until you have enough coverage for all the micro-organisms in it. This means that if you had a large enough dataset, it could be used as a standard for validation.

If you do not have a large enough dataset, however, it is still be true that a larger dataset gives you a higher reliability than a smaller one. This makes it possible to apply the following procedure:

Procedure 2 Subset procedure

1. *Obtain the biggest usable dataset you have.*

2. Apply the error correction to it, and use that as a standard for validation.

3. Take a random subset of it that is one order of magnitude smaller than the whole set.

4. Apply the error correction to the subset and compare with the standard.

Both the full set and the subset came from the sequencing equipment, which means that both have the right sequence distribution and the right error distribution for the test. Since the larger dataset is much more reliable than the smaller one, possibly regardless of the used error correction approach, you can use the difference between the two to measure how different error correction algorithms perform when the available data is too little.

Unfortunately, we were unable to obtain a large enough original dataset to get a significant results. To reliably test the proposed error detection approaches, series of trials will be required.

3 Software solution and conclusion

During the course with our experiments with metagenomic data, it became clear that a flexible yet efficient means to manage and execute procedures will be a very helpful contribution to the bioinformatics community. Any work in the area involves the execution of a mixture of software packages in variety of orders and combinations, which often involves a lot of manual and unnecessary work that is unrelated to the actual study and/or software development.

Just to perform our tests we had to implement two off-the-cuff such programs, one for launching our aligners, and one for performing and managing the error correction. It also became necessary to create a third one, managing the entire workflow in order to perform a more extensive validation, as well as to empirically estimate the parameters of the tested algorithms, while also intelligently storing the intermediate states and results.

At this point we came to the conclusion that it made more sense to create a more general tool that can handle a larger variety of problems. We are now aiming to transform our software to a Python package that:

1. Can execute any required genomic processing task asynchronously so that they can be easily distributed across multiple cores or machines. The interface should be network capable and based on Twisted.

2. Has a simple modular design that allows for easy extensibility, scalability and improvements.

3. Has a common API for every supported task.

4. Access to the functionality through both a Python API, a command-line

interface and a simple task mini-language that allows the execution of simple workflows.

In its current state, the software was written to follow most of these requirements, but it still requires significant clean-up and polishing. A simple, yet flexible package that fulfils this role would be a much more useful contribution to metagenomic studies than any small improvement in error detection qualities, and once it is in a production-ready form we will publish it as free and open source software.

Acknowledgments. This work has been partly funded by the Sofia University SRF within the "Methods and information technologies for ontology building, merging and using" Project, Contract No. 177/2012.

References

1 Jeffreys, A.J., Wilson, V., Thein, S.W. (1984) Hypervariable "minisatellite" regions in human DNA, Nature 314: 67-73.

2 Weisburg, W.G., Barns, S.M., Pelletier, D.A., Lane D., (1991). 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol. 173 (2): 697-703.

3 Kristensen, D.M., Mushegian, A.R, Dolja,V.V., Koonin, E.V. (2009). New dimensions of the virus world discovered through metagenomics. Trends in Microbiology 18 (1): 11-19.

4 Nelson, KE., White, B.A. (2010). Metagenomics and Its Applications to the Study of the Human Microbiome, Metagenomics: Theory, Methods and Applications. 171-182

5 Katoh, K, Kuma, K, Toh, H., Miyata, T. (2005). MAFFT version 5: improvement in accur acy of multiple sequence alignment, Nucleic Acids Research 33 (2): 511-8.

6 Edgar, RC. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5 (1): 113.

7 Zerbino, D.R, Birney, R. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821-829

8 Lu, W., Fu, L., Nui, B., Wooley, J. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis.Briefings in Bioinformatics 6:1-13 doi: 10.1093/bib/bbs035

9 S. Wu, Z. Zhu, L. Fu, B. Niu & W. Li (2011) WebMGA: a Customizable Web Server for Fast Metagenomic Sequence Analysis, BMC Genomics 12:444-452.

10 Toshiaki N., Hachiya T., Tanaka H., Sakakibara Y. (2012) MetaVelvet: an extension of Velvet assembler to denovo metagenome assembly from short sequence reads. Nucleic Acids Research Advance Access published July 19, 2012 doi:10.1093/nar/gks678

Simulation of the Behavior of a Pure Imperative Synchronous Programming Language by Means of Generalized Nets with Stop-Conditions

Magdalina V. Todorova

"St. Kliment Ohridski" University of Sofia, Faculty of Mathematics and Informatics
Sofia 1164, Bulgaria
todorova_magda@hotmail.com

Abstract. The article presents an approach of simulating the behavior of purely imperative synchronous programming languages. For this purpose, generalized nets (GNs) with stop-conditions are used. The simulation is realized for a sample imperative language for synchronous programming which contains the basic programming constructions for purely imperative synchronous programming. After a brief introduction of the language, a GN model is defined for each of its statements. Examples of GN models of programming fragments in the language are given as well.

Keywords: Synchronous Language, Reactive System, Real Time Process, Parallel Programming, Generalized Net, Generalized Net with Stop-Condition

1 Design of Synchronous Language for Reactive Systems

The synchronous programming languages are computer programming languages optimized for programming reactive systems [1]. A reactive system is a system that, when switched on, is able to create desired effects in its environment by enabling, enforcing or preventing events in the environment [2]. The reactive systems form a wide class of systems. These are: real-time systems (safety-critical, embedded, control systems), enterprise resource planning systems, work-flow management systems, groupware systems, e-commerce systems, classic electronic data interchange systems and others.

In this part, we will describe briefly a sample language for synchronous programming whose behavior will be simulated by GNs with stop-conditions [3, 4]. We will call it Synchronous Language for Reactive Systems, abbreviated as *SLRS* for convenience. This language contains the primitive components of known languages of imperative synchronous programming, such as Esterel [5], ECL [6], Jester [7], etc.

The basic object of the language *SLRS* is the *signal*. Signals are used for communication with the environment as well as for internal broadcast communication.

In *SLRS*, there are only two kinds of interface signals: input and output. A *SLRS* program has the following structure:

Program P;
declaration part
interface part
body
end P.

The declaration part

This part declares the external objects used by the program: constants, types, functions, and procedures that manipulate data. They are written in the host language Pascal, C or C++.

The interface part

This part defines the input and output of the program and has the type:

input I_1 {, I_n};
output O_1 {, O_n};
input relations;

where input I_1,..., I_n and output O_1, ..., O_n are the program signals.

Input signals come from the environment. They cannot be produced internally. They are declared in the form:

input I_1 {, I_n};

Output signals are directed towards the environment of the program by the *produce* statement. An output signal declaration has the form:

output O_1 {, O_n};

Input relations are assertions that can be used to restrict input events. They are a very important component of program specification and verification.

A *SLRS* program specifies a relation between input and output signals. It is activated by giving it *input events* repeatedly. These events consist of a possibly empty set of input signals assumed to be present. For each input event, the program reacts by executing its body and by outputting the produced output signals that form the *output event.* We assume that the reaction is perfectly synchronous and deterministic [5]. Deterministic reactive program produces identical output sequences when fed with identical input sequences [5].

The part "body"

The *body* is an executable statement. The statements in the language are:

Statement skip

skip

It performs no action and terminates immediately.

Statement stop

stop

It performs no action and never terminates.

Statement produce

produce S

where *S* is a signal.

It emits *S* and terminates immediately.

Statement sequence

sequence $stat_1$, $stat_2$ end

where $stat_1$ and $stat_2$ are any statements.

The statement $stat_2$ starts instantly when the statement $stat_1$ terminates. The sequencing operator takes no time by itself.

Statement parallel

parallel $stat_1$, $stat_2$ end

where $stat_1$ and $stat_2$ are any statements.

The statements $stat_1$ and $stat_2$ are started simultaneously when the parallel statement is started. The *parallel* statement terminates when both of its branches are terminated.

Statement ifp-then-else

ifp S then $stat_1$ else $stat_2$ end

where *S* is a signal, $stat_1$ and $stat_2$ are any statements. The *then* and *else* parts are optional. If some of them is omitted, it is assumed to be a *skip* statement.

The presence of *S* is tested and the *then* or *else* branches start immediately, respectively.

Statement cycled-end

cycled stat end

where *stat* is any statement.

The body *stat* of a *cycled-end* statement starts immediately when the *cycled-end* statement starts and whenever *stat* terminates, *stat* is instantly restarted. A *cycled-end* never terminates.

Statement watching-do

watching S do stat end

where *stat* is any statement and *S* is a signal. *S* is called *a guard*.

The statement *stat* is executed normally until *stat* terminates or until further occurrence of the signal *S*. If *stat* terminates just before *S* occurs, so does the whole

watching-do statement and the guard takes no action. Otherwise, the occurrence of *S* provokes immediate preemption of the body *stat* and immediate termination of the whole *watching-do* statement.

Statement run-until

run stat until X

where *stat* is any statement and *X* is a parameter.

The body *stat* starts instantly and determines the behavior of the *run-until* statement until it terminates or executes *exit X*. Then the execution of *stat* is preempted and the whole *run-until* constructor terminates. If body of a *run-until* statement contains parallel components the *run-until* is exited when one of the components executes an *exit X*, the other component is preempted.

Statement local

local S {, S_i} in stat end

where *S* and S_i are signals and *stat* is any statement.

It declares the scope of the signal *S* {, S_i} that can be used for internal broadcast communication within *stat.*

At each reaction, the signal has a single status - *present* or *absent.* The following law determines the status of local and output signals: A local or output signal is present in a reaction if and only if it is produced by executing a *produce* statement in that reaction. The default status of a signal is absent.

2 GN-models of the *SLRS* programs

GNs with stop-conditions are used to realize models of *SLRS* language programs. This choice is motivated by:

a. By means of GNs, it is possible to create the most detailed and accurately modeled real-time processes. This is confirmed by many models which are made by such means.

b. GNs are well-studied from a mathematical point of view [3] and [4]. Their theory is enriched with algebraic, topological, logical, operator and methodological aspects.

c. By means of GNs, it is possible to verify properties of real-time processes.

d. By means of GNs, it is possible to seek optimal ways to conduct real-time processes.

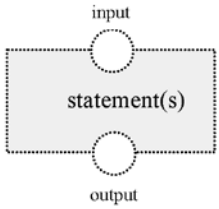
e. By means of GNs, it is possible to search for ways to improve real-time processes.

f. Automated tools for modeling GNs have been developed. They are united

in the environment GN Lite [8, 9, 10 and 11]. Through them, experiments may be carried out to verify the established models of the synchronous programs.

g. GNs are easy to apply by software engineers.

The body of the program in the language *SLRS* is created by connecting appropriately its language constructions. In order to create a GN-model of the program, each language construct of *SLRS* is transformed in a corresponding *GN*. Joining them in a statement requires GNs corresponding to the linguistic structures to be constructed in a way that allows GNs integration. This is achieved through building the GN-models of the compound language using the same pattern. The pattern of the proposed approach has the following form:



where *input* denotes one or more input places and *output* – one or more output places of the GN-pattern. The relationship between the GN-patterns is realized through the places *input* and *output*. The latter forms the interface of the pattern.

A pair of the type <*E*, *P*> is chosen to characterize the token of the *GN* which simulates the behavior of an operator or a program unit. Here *E* is a set containing the available signals, and *P* is a set of parameters. The parameters are identifiers that are involved in the *run-until* and *exit* operators. When executing the transition which starts the realizing the operator *run stat until X* net, a parameter *X* is added to the set of parameters *P*. Similarly, executing the transition realization of the operator *exit X* causes excluding of the parameter *X* of the set of parameters *P*.

There are operators of the language which do not complete their execution; therefore we will choose a *GN* with stop-conditions for modeling them. The possible cases are:

a) the operator is not compound

If the operator completes its execution, the stopping condition of the modeling transition is *false*. If the operator does not complete its execution, the stop-condition of the modeling transition is a condition representing interrupting the execution of a transition because of activating a particular signal or depends on time (if the current value of the modeling time for executing the transition $t_{current}$ exceeds the predetermined maximum execution time t_{max}).

b) the operator is compound

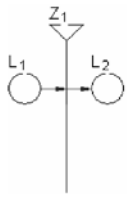
If any of the operators of the compound operator does not complete the execution, the compound operator will not complete execution as well. The *GN*

which simulates the compound operator behavior has a stop-condition which is a disjunction of the stop-conditions of all the operators it comprises. Therefore, in case any of the operators within the compound one does not complete its execution, the *GN* which represents the compound operator stops its execution as it stop-condition receives the value *true*.

2.1 GN-models of the *SLRS* statements

In order to facilitate the presentation, some details are not given in the following description. The stop-condition of the transition Z we will denote C_Z, and C_{stat} is the stop-condition of the operator *stat*.

GN-model of statement skip

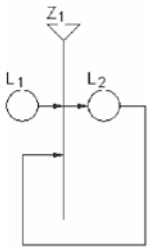


$$Z_1 = \langle \{L_1\}, \{L_2\}, r_1, C_{Z_1} \rangle$$

$$r_1 = \begin{array}{c|c} & L_2 \\ \hline L_1 & \text{true} \end{array} \; ; \; C_{Z_1} = \text{false}$$

The characteristic of the token in the place L_1 does not change when it is transfered in a place L_2.

GN-model of statement stop

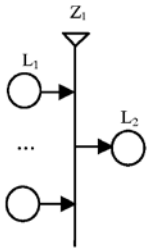


$$Z_1 = \langle \{L_1, L_2\}, \{L_2\}, r_1, C_{Z_1} \rangle$$

$$r_1 = \begin{array}{c|c} & L_2 \\ \hline L_1 & \text{true} \\ L_2 & \text{true} \end{array} \qquad C_{Z_1} = t_{current} > t_{max}$$

The characteristic of the token in the place L_1 does not change when it is transfered in a place L_2.

GN-model of statement produce S



$$Z_1 = \langle \{L_1, \ldots\}, \{L_2\}, r_1, C_{Z_1} \rangle$$

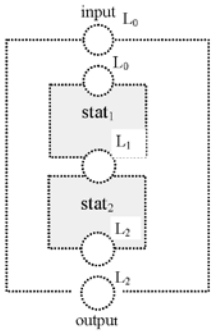
$$r_1 = \begin{array}{c|c} & L_2 \\ \hline L_1 & \text{true} \\ \ldots & \text{true} \end{array} \qquad C_{Z_1} = \text{false}$$

If the characteristic of the token in place L_1 is <E, P>, it gets the type <E ∪ S, P> when the token moves to place L_2.

The descriptions of the *GN* models of the operators *sequence*, *parallel*, *ifp*, *cycled* and *watching* given below are valid for the case when these operators do not contain the operator *exit*. The existence of the operator *exit* changes the *GN* architecture, as well as the parameter *P* of the token characteristic. The third example in 2.2 illustrates this.

GN-model of statement sequence $stat_1$, $stat_2$ end

Let the characteristic of the token in the place L_0 (see below) is <E, P>. The stop-condition of the net is $C_{sequence} = C_{stat1}$ or C_{stat2}



a) $C_{sequence}$ = false

When the token reaches position L_1, its characteristic becomes $<E \cup O_1, P>$, where O_1 denotes the set of output signals, resulting from the execution of $stat_1$. Respectively, when the token reaches position L_2, its characteristic becomes $<E \cup O_1 \cup O_2, P>$, where O_2 denotes the set of output signals, resulting from the execution of $stat_2$.

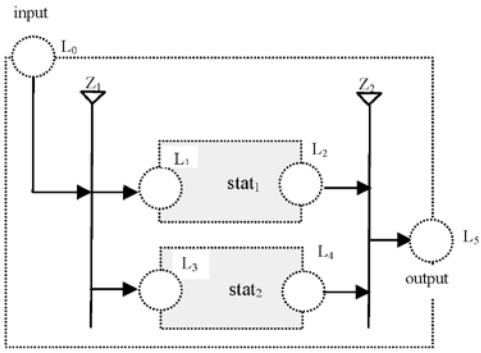
b) C_{stat1} = false, C_{stat2} = true

In this case the execution of the *GN* of the operator *sequence* ends when its stop-condition becomes *true*. If $<E \cup O_1, P>$ is the characteristic of the token in the place L_1, in the place L_2 the characteristic becomes $<E \cup O_1 \cup O_2, P>$. O_2 denotes the set of output signals, resulting from the $stat_2$ execution up to the moment of its complete execution under the stop-condition.

c) C_{stat1} = true

In this case the execution of the *GN* of the operator *sequence* ends when the execution of the net of the statement $stat_1$ ends.

GN-model of statement parallel $stat_1$, $stat_2$ end



$Z_1 = \langle \{L_0\}, \{L_1, L_3\}, r_1, C_{Z_1} \rangle$

$r_1 =$

	L_1	L_3
L_0	true	true

$Z_2 = \langle \{L_2, L_4\}, \{L_5\}, r_2, C_{Z_2} \rangle$

$r_2 =$

	L_5
L_2	true
L_4	true

$C_{Z_1} = C_{Z_2}$ = false

$C_{parallel} = C_{stat1}$ or C_{stat2}

Let the characteristic of the token in place L_0 is <E, P>. The transition Z_1 is executed unconditionally and causes splitting of the token located in a place L_0 of two tokens that move in the places L_1 and L_3 without changing the characteristic.

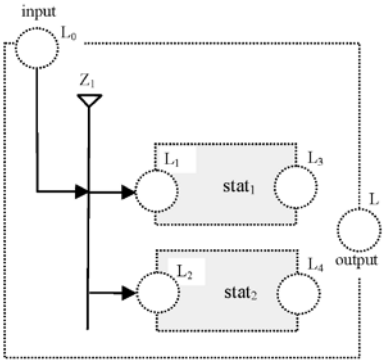
a) $C_{parallel}$ = false

After execution of $stat_1$ the token in place L_1 moves to place L_2, and its characteristic becomes $<E \cup O_1, P>$, where O_1 denotes the set of output signals which are received as a result of performing $stat_1$. Respectively, after execution of $stat_2$, the token in place L_3 is transferred to place L_4, and its characteristic becomes $<E \cup O_2, P>$, where O_2 denotes the set of output signals received as a result of performing $stat_2$. After execution of the transition Z_2 the tokens in L_2 and L_4 merge in a token with characteristic $<E \cup O_1 \cup O_2, P>$ in place L_5.

b) C_{stat1} or C_{stat2} = true

In this case, the operators $stat_1$ or $stat_2$ (or both) do not terminate. Their GN-models end the execution when stop-conditions are fulfilled.

GN-model of statement ifp S then $stat_1$ else $stat_2$



$Z_1 = \langle \{L_0\}, \{L_1, L_2\}, r_1, C_{Z_1} \rangle$

$$r_1 = \begin{array}{c|cc} & L_1 & L_2 \\ \hline L_0 & S \in E & S \notin E \end{array}$$

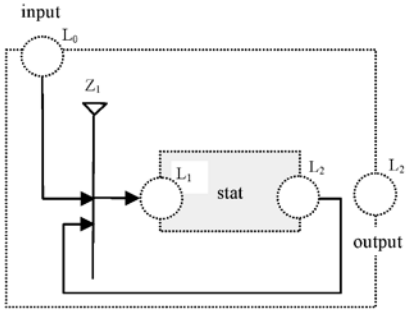
C_{Z_1} = false

L denotes the output positions of $stat_1$ and $stat_2$.

$$C_{ifp_then_else} = \begin{cases} C_{stat1}, & \text{ако} S \in E \\ C_{stat2}, & \text{ако} S \notin E \end{cases}$$

Let the token characteristics in L_0 is <E, P>. After completing the execution of the GN model, its token characteristic will be either $<E \cup O_1, P>$ or $<E \cup O_2, P>$, depending on whether the signal *S* belongs or does not belong to the set of signals *E*. O_1 and O_2 denote the sets of output signals, resulting from the execution of the *GN* of the operators $stat_1$ and $stat_2$, respectively.

GN-model of statement cycled stat end



$Z_1 = \langle\{L_0, L_2\}, \{L_1\}, r_1, C_{Z_1}\rangle$

$C_{Z_1} = t_{current} > t_{max}$

$r_1 =$

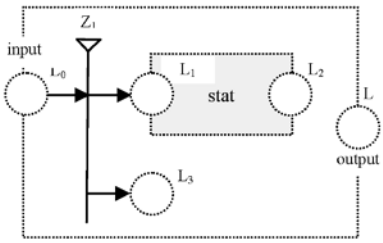
	L_1
L_0	true
L_2	true

$C_{cycled} = C_{Z_1}$ or C_{stat}

The token characteristic, which in L_0 is <E, P>, after completing the *GN* model under the stop-condition becomes <E ∪ O, P>. Here O = (i = 1, 2, …). O_i is the set of output signals, received as a result of *i* sequential executions of the operator *stat* (*i* depends on the stop-condition of the net).

GN-model of statement watching S do stat end

If the signal *S* is activated during the execution of the operator *stat*, the execution of the net, which realizes *stat*, is interrupted. The active tokens in it are deactivated. A new token is activated in an output position *L*. It has a characteristic < E ∪ O, P>, where *O* is the set of output signals received during the execution of the operator *stat GN*.



$Z_1 = \langle\{L_0\}, \{L_1, L_3\}, r_1, C_{Z_1}\rangle$

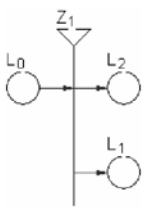
$r_1 =$

	L_1	L_3
L_0	$S \notin E$	$S \in E$

C_{Z_1} = false; $C_{watching} = C_{stat}$

L denotes the output positions L_2 and L_3.

GN-model of statement exit X



$Z_1 = \langle\{L_0\}, \{L_1, L_2\}, r_1, C_{Z_1}\rangle$

$r_1 =$

	L_1	L_2
L_0	$X \notin P$	$X \in P$

; C_{Z_1} = false

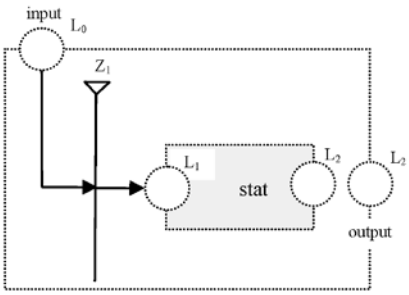
Position L_2 coincide with the output position of the operator *run stat until X* corresponding to *exit X*.

Let the characteristic of the token in L_0 is <E, P>. If the condition holds, the program has an error (lack of correspondence between the operators *run-until* and *exit*). Otherwise, *exit X* performs a destructive action on all tokens in the positions

of the *GN* which realize the operator *stat*: operator body *run stat until X*. A new token is activated in output position L_2 of *run stat until X,* characterized by <E ∪ O, P \ X>, where O = (i = 1, 2, …), and O_i are the set of output signals, received as a results of executing the operator *stat* of *run-until*.

GN-model of statement run stat until X

If the token characteristic in L_0 (see below) is <E, P>, when the token moves to position L_1, the characteristic changes into <E, P ∪ X>. If the operator *stat* completes execution (without triggering the operator *exit X*), the token transfers to position L_2 and receives a characteristic of the type <E ∪ O', P ∪ X>. Here *O'* is the set of the received output signals after executing *stat*.



$Z_1 = \langle \{L_0\}, \{L_1\}, r_1, C_{Z_1} \rangle$

$r_1 =$

	L_1
L_0	true

C_{Z_1} = false

$C_{run(X)} = C_{stat}$

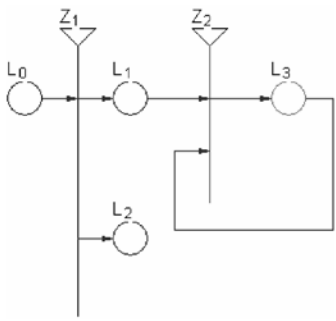
If the operator *exit X* is executed, the execution of the operator *run stat until X* ends. The token characteristic in position L_2 in this case is of the same type as offered in the description of the model of the operator *exit X*.

2.2 Some examples

The GN-model of the operator

await S = waiting S do stop end

has the type



$Z_1 = \langle \{L_0\}, \{L_1, L_2\}, r_1, C_{Z_1} \rangle$

$r_1 =$

	L_1	L_2
L_0	$S \notin E$	$S \in E$

$Z_2 = \langle \{L_1, L_3\}, \{L_3\}, r_2, C_{Z_2} \rangle$

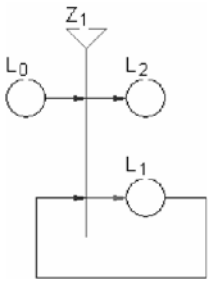
$r_2 =$

	L_3
L_1	true
L_3	true

C_{Z_1} = false; C_{Z_2} = S is activated.

or $t_{current} > t_{max}$

This net can be simplified to the following net:



$Z_1 = \langle \{L_0, L_1\}, \{L_1, L_2\}, r_1, C_{Z_1} \rangle$

$r_1 =$

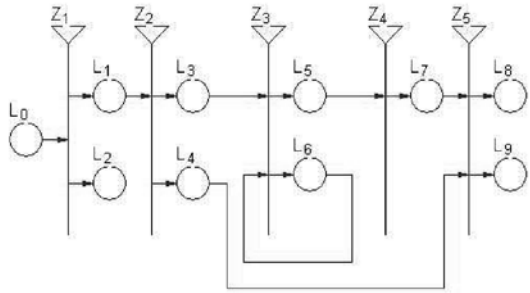
	L_1	L_2
L_0	$S \notin E$	$S \in E$
L_1	$S \notin E$	$S \in E$

$C_{Z_1} = t_{current} > t_{max}$

This simplified net is applied as the operator *await* is often used.

A program fragment in *SLRS* language and its *GN* model are to follow:

```
watching I1 do
   sequence
      watching I2 do
         sequence
            await I3,
            produce O1
         end
      end,
      produce O2
   end
end
```



Here

$Z_1 = \langle \{L_0\}, \{L_1, L_2\}, r_1, C_{Z_1} \rangle$

$r_1 =$

	L_1	L_2
L_0	$I_1 \notin E$	$I_1 \in E$

$Z_2 = \langle \{L_1\}, \{L_3, L_4\}, r_2, C_{Z_2} \rangle$

$r_2 =$

	L_3	L_4
L_1	$I_2 \notin E$	$I_2 \in E$

$Z_3 = \langle \{L_3, L_6\}, \{L_5, L_6\}, r_3, C_{Z_3} \rangle$

$r_3 =$

	L_5	L_6
L_3	$I_3 \in E$	$I_3 \notin E$
L_6	$I_3 \in E$	$I_3 \notin E$

$Z_4 = \langle \{L_5\}, \{L_7\}, r_4, C_{Z_4} \rangle$

$r_4 =$

	L_7
L_5	true

$Z_5 = \langle \{L_4, L_7\}, \{L_8, L_9\}, r_5, C_{Z_5} \rangle$

$r_5 =$

	L_8	L_9
L_4	false	true
L_7	true	false

$C_{Z_1} = C_{Z_2} = C_{Z_4} = C_{Z_5} = \text{false};$

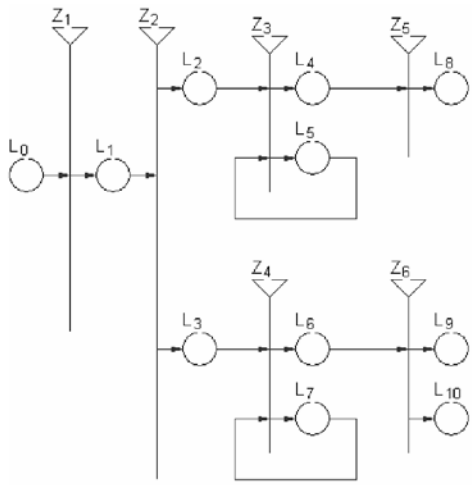
$C_{Z_3} = t_{current} > t_{max}$

In this example, if <E, P> is the characteristic of the token in the place L_0, the set of signals *E* becomes $E \cup O_1$ in the place L_7, and it becomes $E \cup O_2$ in position L_8. Here are realized the following actions: if *I1* occurs before *I2* and

I3, or at the same time as them, then the external *watching-do* preempts its body and terminates instantly. In this case, no signal is produced. If *I2* occurs before or simultaneously with *I3*, but before *I1*, then the internal *watching* preempts its body, *O1* is not produced even if *I3* is present, *O2* is produced and the external *watching* instantly terminates. In case *I3* occurs just before *I1* and *I2*, then the *await* statement terminates, *O1* is produced, the internal *watching-do* terminates since its body terminates, *O2* is produced and the external watching also terminates.

One more example for another program fragment in *SLRS* language and its *GN* model:

```
run
  parallel
    sequence
      await I1,
      produce O
    end,
    sequence
      await I2,
      exit X
    end
  end
until X
```



Here

$Z_1 = \langle \{L_0\}, \{L_1\}, r_1, C_{Z_1} \rangle$

$$r_1 = \begin{array}{c|c} & L_1 \\ \hline L_0 & \text{true} \end{array}$$

$Z_2 = \langle \{L_1\}, \{L_2, L_3\}, r_2, C_{Z_2} \rangle$

$$r_2 = \begin{array}{c|cc} & L_2 & L_3 \\ \hline L_1 & \text{true} & \text{true} \end{array}$$

$Z_3 = \langle \{L_2, L_5\}, \{L_4, L_5\}, r_3, C_{Z_3} \rangle$

$$r_3 = \begin{array}{c|cc} & L_4 & L_5 \\ \hline L_2 & I_1 \in E & I_1 \notin E \\ L_5 & I_1 \in E & I_1 \notin E \end{array}$$

$Z_4 = \langle \{L_3, L_7\}, \{L_6, L_7\}, r_4, C_{Z_4} \rangle$

$$r_4 = \begin{array}{c|cc} & L_6 & L_7 \\ \hline L_3 & I_2 \in E & I_2 \notin E \\ L_7 & I_2 \in E & I_2 \notin E \end{array}$$

$Z_5 = \langle \{L_4\}, \{L_8\}, r_5, C_{Z_5} \rangle$

$$r_5 = \begin{array}{c|c} & L_8 \\ \hline L_4 & \text{true} \end{array}$$

$Z_6 = \langle \{L_6\}, \{L_9, L_{10}\}, r_6, C_{Z_6} \rangle$

$$r_6 = \begin{array}{c|cc} & L_9 & L_{10} \\ \hline L_6 & X \in P & X \notin P \end{array}$$

$C_{Z_1} = C_{Z_2} = C_{Z_5} = C_{Z_6} = \text{false}$

$C_{Z_3} = C_{Z_4} = t_{current} > t_{max}$

In this example, if <E, P> is the characteristic of the token in place L_0,

the characteristic becomes <E, P ∪ X> in place L_1. The set of signals *E* of the characteristic changes in place L_8 and becomes E ∪ O if the execution reaches place L_8. The execution of the transition Z_6 ends the execution of the statement *run-until X*. As a result, all tokens in the *GN* are deactivated and a token with a characteristic <E′,P> is activated in place L_9, where *E′* includes all signals in the *GN*.

In the program part, the following actions are realized: if *I1* occurs before *I2*, then *O* is produced and run waits for *I2* to terminate; if *I2* occurs before *I1*, then the whole statement terminates instantly, the first branch is preempted and *O* will never be produced; if *I1* and *I2* occur simultaneously, then both branches do execute and *O* is produced.

3 Conclusion

The article presents new means for simulating the behavior of purely synchronous imperative programs. In order to achieve this, GNs with stop-conditions are used as modeling apparatus. By applying the designed models and with the help of the *GN Lite* software system, we can simulate the behavior of reactive systems; verify the GN-defined models of reactive systems to perform optimization of the realization. The idea described above is in an initial stage of development. Designing and researching models of particular reactive systems and selecting of appropriate verification strategies are to follow [12, 13]. Analysis of the applicability of the designed model is also under consideration, as well as comparing it to other models.

Acknowledgements. The work has been supported by the Sofia University Research Fund within Project 127/2012.

References

1. http://en.wikipedia.org/wiki/Synchronous_programming_language (last visited in May 2012).
2. http://www.elsevierdirect.com/companions/9781558607552/slides/slides.pdf (last visited in May 2012).
3. Atanassov, K.: On Generalized Nets Theory, Prof. Marin Drinov Academic Publishing House, Sofia, 2007.
4. Atanassov, K.: Generalized Nets, World Scientific, Singapore, 1991.
5. Berry, G., Gonthier, G.: The Esterel synchronous programming language: Design, semantics, implementation. Science of Computer Programming, 19, 2 (1992), 87–152.
6. http://ecl.sourceforge.net/index.htm#top (last visited in May 2012).
7. http://www.cfdvs.iitb.ac.in/download/Docs/verification/tools/jester/html/roadmap/node7.html (last visited in May 2012).

8. Trifonov, T., Georgiev, K.: GNTicker – A Software Tool for Efficient Interpretation of Generalized Net Models. Issues in Intuitionistic Fuzzy Sets and Generalized Nets, Vol. 3. Warsaw, 2005.
9. Trifonov, T., Georgiev, K., Atanassov, K.: Software for Modelling with Generalized Nets. Issues in Intuitionistic Fuzzy Sets and Generalized Nets, Vol. 6, 2008, 36-42.
10. Dimitrov, D. G.: Optimized Algorithm for Token Transfer in Generalized Nets, Proc. of 9th IWIFSGN 2010, 8 October 2010, Warsaw, Poland.
11. Atanassov, K., Dimitrov, D., Atanassova, V.: Algorithms for Tokens Transfer in the Different Types of Intuitionistic Fuzzy Generalized Nets. Cybernetics and Information Technologies, Vol. 10, 2010, No. 4, 22-35.
12. Kaloyanova, K., Ignatova, P.: Software Testing Automation, Second International Scien-tific Conference"COMPUTER SCIENCE 2005", Chalkidiki, Greece, September, 2005, pp 220-225.
13. Maneva, N., Grozev, N., Lilov, D.: A Framework for Source Code Metrics. Proc. of Comp SysTech'2010, Sofia, 17-18 June, 2010, pp.113-118.

Declarative Semantics of the Program Loops*

Krassimir Manev and Trifon Trifonov

Faculty of Mathematics and Informatics, Sofia University

ACM Classification Codes: D.2.7, D.2.5.

Keywords: Business Rules, Extraction of BR from Source Code, Static Analysis, Loops, Declarative semantics.

Abstract. The business rules (BR) approach has been introduced at the end of the past century with the goal to facilitate the specification of business software and to make it more adequate to the needs of the corresponding business. Nowadays most of the stated goals of the approach have been achieved. But the efforts for providing "... a rigorous basis for reverse engineering BR from existing systems..." are still in progress. In a previous paper we described an approach for deriving BR from source code, based on methods for source code static analysis. One of the main problems of the applying the static analysis approach for deriving rules from procedural program code is that the "language" of the BR is **declarative**. Therefore, for the successful extraction of BR from procedural programming code, their iterative (procedural) semantics has to be replaced by the declarative equivalent. The paper identifies some patterns of programming loops in the source code and shows how a declarative semantics, corresponding to each of these patterns, could be defined.

1 Introduction

The rapid development of the software technologies in the end of the 20th and the beginning of the 21st century the phenomenon called *legacy systems* (LS) was born. In [Bisbal et al., 1999] the following four characteristics of a LS are outlined (the authors of the paper mean Legacy *Information Systems*, but the observations are valid for an arbitrary computerized system):

- LSs usually run on legacy hardware — slow and expensive to maintain;
- Maintenance of LSs can also be expensive, due to lack of documentation and understanding of detail;

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

- Lack of clear interfaces makes integrating and inter-operability of LSs with other systems very difficult;
- Extending of LSs is also very difficult, if not impossible.

Initial semantics of the word *legacy* in *legacy system* is negative. It seemed that the simplest possible solution of the problem is to stop usage of legacy software and to replace it by new one built from scratch. But the practice is different.

It is supposed that in 2006 about 60-70% of the working corporative software was written in COBOL — a language that is not maintained in the moment, not included in university curricula, and so — unknown for the nowadays software engineers. That means that in some enterprisers all staff don't know other software except the legacy corporative system written in COBOL. And these systems are stable, well-tested and well-adapted to the needs of the business.

That is why [McGee, 2005] asserts: "The time when we could think of "legacy" as a pejorative term within the technology field is gone. As more and more of our technology efforts connect to (and integrate with) existing systems both inside and outside our organizations, the more we must cope with legacy systems." The intelligent way to resolve the problem is to transform smartly the legacy systems that proved their qualities to a modern platform using them as specifications of themselves.

One possibility for intelligent transformation of legacy systems is to use the Business Rules approach which became more and more popular recently. The Business Rules Project (BRP or GUIDE BRP) started with an aim "to formalize an approach for identifying and articulating the rules which define the structure and control the operation of an enterprize". In its Final report [Hay&Healy, 2000], members of the research team outlined four main objectives of the project, and among them:

- To provide a rigorous basis for reverse engineering of business rules from existing systems,

i.e., the authors of the BR concept predicted that the approach could be used for reverse engineering of the existing systems. Starting from the model of IBM [Stineman, 2009] for BR-based development, an initial version of a model for *BR-based modernization of a legacy system* is proposed in [Maneva&Manev, 2011]. One of the most important stages in our model is extraction the built-in BR from the legacy system.

As it was concluded in [Maneva&Manev, 2011], one of the most valuable for the extraction of BR resources of the legacy system is the *source*

code of the software part of the system. That is why in [Manev et al., 2012] an attempt was made to construct an algorithm, which split the code to pieces, appropriate for building back the business logic embedded in the code. The algorithm is based on a *static analysis* used by authors in other projects [SSA, 2012].

This paper is dedicated to a problem that arises in a process of splitting of the code to BR-like pieces. By definition, BR are expressed in **declarative form** and so the iterative parts of the code — i.e., loops — could not be transformed in declarative BR-like pieces directly. In Chapter 2 the necessary notions will be defined and a more formal definition of the tasks will be given. Chapter 3 contains the essential part of the work — an attempt for elimination of the procedural semantics of loops through *re-functionalization* of the iterative part of code. Some conclusion and perspectives for future research are given in Section 4.

2 Preliminaries

In this chapter we will present the main notions that are necessary for presenting the solved task as well as one example to illustrate the conceptions.

2.1 Business rules

Following [Hay&Healy, 2000], a business rule is "a statement that defines or constrains some aspect of the business. It is intended to assert business structure or to control or influence the behavior of the business." Business rules are classified in following four kinds:

- *Business terms* are the elementary business rules (*words*) that are used to build the other business rules. They are usually documented in glossaries (or as entities in entity/relationship diagrams).
- *Facts* relate terms to each other. The structure of an organization can be described with the facts which relate terms. Facts can be documented as natural language sentences or as relationships.
- *Constraints* permit or prevent an action to be taken. Each enterprise constrains its behavior in some way.
- *Derivations* define how knowledge in one form have to be transformed into other knowledge, possibly in a different form.

2.2 Sample code

For the consideration below we will use the following sample of programming code — simplified version of a real function:

```
public VacReqResult reqVac(Employee anEmployee,          //1
       Date fromDate, Date toDate, VacType type)         //2
{  int daysLeft = anEmployee.getVacDays(type);           //3
   int daysReq = CompCalend.getWorkDays(fromDate,toDate);//4
   if (daysLeft < daysReq)                               //5
   {   return VacReqResult.AUTO_REJECT;  }               //6
   else                                                  //7
   {   Employee operLeader = anEmployee.getOperLeader(); //8
       while(operLeader.isOnVac())                       //9
       {   operLeader = operLeader.getOperLeader(); }    //10
       notifyForVacReq(operLeader);                      //11
       return VacReqResult.REQUESTED;                    //12
   }                                                     //13
}                                                        //14
```

Remark. For the purposes of the paper the source code of the sample is too nice (verbose) — all identifiers clearly show the purpose of corresponding type, variable, constant or function/method. In the reality this will not be the case.

2.3 Extraction of BR through static analysis

Static analysis is performed over the source code of the program without executing it. [Manev et al., 2012] used static analysis approach with *covering the paths* in the *control flow graph* of the program module under analysis. This approach is well known for building automated test data generators, for example. The steps of the approach with covering of the paths are displayed in Figure 1.

◇ *Parsing the program.* At this stage the code is parsed and the algorithm build the simple rules — terms and facts — from the declarative parts of the code: terms from variable names, constants, function names, names of the user defined types; facts from declarations that associates a variable with its type, definitions of aggregated types, definitions of user types, definitions of functions or function calls, and so on. Then a control flow graph of the program (CFG) and a *data dependence graph* (DDG) for each local variable are constructed. DDG of a variable is used to trace

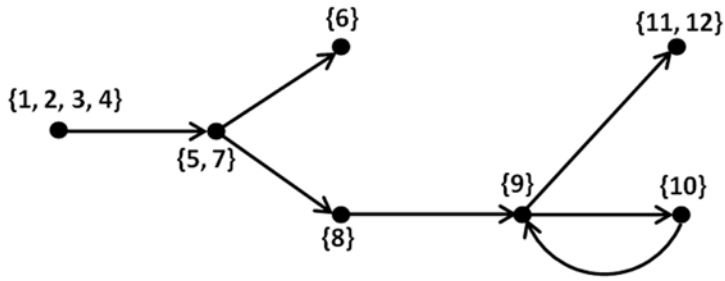


Fig. 1. Control Flow Graph

changes of its value during the execution of the program. The CFG for the sample program is shown on the Figure. Each vertex of CFG is labeled with sequences of non branching operators or with a single branching operator. The CFG of each program module is directed with one minimal vertex (the entrance point) and one or more maximal vertices (end points).

◇ *Path exploring.* At this stage a set of paths is chosen so as to "cover" the CFG of the program. Each path start in the minimal vertex and finish in some of the final vertices. The CFG of our example is covered by two paths:

```
{1,2,3,4}, {5,7}, {6}
```

and

```
{1,2,3,4}, {5,7}, {8}, {9], ({10}, {9])*, {11,12}
```

where is the Kleene star representing the repetitions of the loop's body.

◇ *Extracting the rules.* This stage is composed of three sub-stages. During this first sub-stage, each vertex labeled with sequence of not branching operators is transformed in *derivation element.* DDG of the involved variables are used to describe changes in the variables produced by the derivation element. During the second substage each vertex labeled with single branching operators is transformed in a *constraining element.* The final third substage of the algorithm composes one candidate rule for each covering path identified at the previous stage.

2.4 The problem

The problem is that the paths containing loops could not be transformed in BR because the proclaimed in [Hay&Healy, 2000] principle that the

BR could have only declarative form. That is why we will try in the rest of the paper to approach this problem and to propose a solution to the problem.

3 Functionalization of iterative semantics

3.1 Declarative semantics of procedural programs

Procedural programming languages are heavily based on the computational model of a Turing machine, in which a program is defined as a finite sequence of instructions, performing atomic operations on the computer memory. The underlying tool for describing non-trivial computational behavior is the *loop*, or an iteration of a list of instructions while a given condition of the machine state evaluates to true. The natural semantics of languages with iteration are the operational semantics, which define the behaviour of each atomic instruction and specify how a sequence of instructions is translated to a composition of the semantics. The semantics of sequential programs rely on the notion of a "machine state". One of the difficulties with the extraction a business rule from a given procedural program is the necessity to abstract away from the state, which defines the context in which given operations are executed. While in some cases, the context of a given operation can be very small and easy to handle, there are programs whose logic is quite entangled and depends on many local variables. The complexity of the context in such cases would become too large for a sensible business rule to be extracted.

One of the notable cases of non-operational semantics of iterative programs are Dijkstra's axiomatic "weakest precondition" semantics:

```
{Precondition} Operator {Postcondition}
```

In such a case, a rule like the following could be extracted:
Whenever ***Precondition*** *holds,* ***Postcondition*** *should hold after* ***Operator*** *is applied.*

However, when a variable value is modified in the operator, and this value is referred to both in the precondition and postcondition, suitable variable renaming needs to take place, so that the rule reflects the presence of a modifiable state of the program. A loop is an extreme example of this form, as there exist *loop variables* appearing in the body of the loop as well as the condition of the loop. We thus obtain for each loop variable a trace of all possible values of a given loop variable, and BR extraction becomes close to impossible. One way to circumvent this problem is the

concept of a *loop invariant*, i.e., a condition which holds true throughout the whole loop:

```
{Precondition}
while(Condition)
  {Invariant}
  Operator
{Postcondition}
```

In such cases, the following business rules could be extracted:
Whenever ***Precondition*** *holds and* ***Condition*** *holds, then* ***Invariant*** *should also hold.*
Whenever ***Condition*** *holds and* ***Invariant*** *holds, then* ***Invariant*** *should also hold after* ***Operator*** *is applied.*
Whenever ***Condition*** *does not hold,* ***Postcondition*** *holds.*

The major problem of using such axiomatic semantics to generate BR is the necessity to reflect changes to the state of the program, expressed in the examples above with the words ***Operator*** *is applied.* While making the step towards being more expressive, **Operator** might be so complicated, and the corresponding business rule could become as long as the program itself.

On the other hand, the natural semantics of functional programming languages are declarative. Being free from side effects and not relying on a state, functional programs describe the features of the computation, as opposed to its steps. Using λ-calculus as a computational model, functional programming languages tend to be very expressive, with recursive constructs being at the core of the syntax. In the late 40s of the 20th century it has been shown that the Turing machine and λ-calculus models are expressively equivalent. Therefore, each iterative program could be expressed in a functional programming language. The following subsection explores different ways to perform this formal transformation, which helps unfolding the declarative semantics of the original iterative program.

3.2 Functionalization of procedural programs

It has been shown that iteration can be viewed as a special case of recursion. Consider the following snippet from the program above:

```
Employee operLeader = anEmployee.getOperLeader(); //8
while(operLeader.isOnVac())                       //9
{   operLeader = operLeader.getOperLeader(); }    //10
```

A recursive way to express this is as follows:

```
boolean operLeader(anEmployee) {
    if (anEmployee.getOperLeader().isOnVac())
    { return operLeader(anEmployee.getOperLeader()); }
    else return anEmployee.getOperLeader();
}
```

The above recursion pattern is usually referred to as "tail recursion", as the recursive call does not introduce delayed operations and thus no backtracking is necessary. Scheme is one of the first languages to optimize the implementation of tail recursion, by translating it to iteration. Note that we can now formulate a clear BR from this functional program:
Whenever the operational leader of an employee is on vacation, their operational leader becomes the new operational leader of the employee.

One could argue that we could have extracted the same business rule from the iterative program above, by reformulating the procedureal "while" keyword as a declarative "if" statement. This is indeed the case; functionalizing simple iterative programs often does not bring any advantages in the light of understanding the business logic behind the program. However, as an iteration involves more variables, the function parameters of the functionalized program increases, and so does the complexity of formulating a BR.

Fortunately, functional programming has another dimension of expressiveness which can deal with handling of complicated states. By treating functions as first-class citizens of the language, functions can be generated, passed as parameters and returned as results. Using such higher-level functions, one could significantly reduce the size of a program, thus improving its expressiveness.

The continuitiy of computable functions implies that only a finite part of any function passed as a parameter is needed for any given computation. Therefore, functional parameters could be approximated by a restrictng their domain on an appropriate finite range of values. In its turn, finite functions can be expressed as atomic types, which provides us a way to transform a functional program of a higher order to an iterative program. Such an approach has been explored by Danvy and Nielsen in [Danvy&Nielsen, 2001].

However, in our case, it is the reverse direction of the transformation which is more relevant, i.e., obtaining a higher order functional program from an iterative program. This inverse transformation has been explored by Danvy and Millikin in [Danvy&Millikin, 2009] and is referred to as "refunctionalization". The following is a reformulation of an example, given as a use case of the refunctionalization technique.

Consider an ongoing dialogue between two communication nodes, in which the following messages are allowed:

- `REQ` — a request
- `RESP` — a response
- `STOP` — denotes the end of the conversation

Let us consider a program which recognizes whether in a given dialogue every request is eventually given exactly one response. This means that requests and responses are in a bijective correspondence as follows:

- for every request there is a matching following response, and
- for every response there is a matching preceding request.

The following program determines whether a given process is valid, by utilizing a counter for pending requests.

```
Message m = conversation.getFirst();
int count = 0;
while(!m.equals("STOP"))
{
  if (m.equals("REQ"))
    count++;
  if (m.equals("RESP"))
    count--;
  if (count < 0)
    return false;
  m = conversation.dropFirst().getFirst();
}
if (count > 0)
  return false;
return true;
```

One can easily see that it is not straightforward to formulate a business rule corresponding to the program above solely based on its syntactic form. The complication comes from the semantics of the variable `count`; its value serves as an assertion for the correction of the bijection.

Applying a formal syntactic refunctionalization transformation, Danvy and Millikin obtain a functional program similar to the one below:

```
boolean check(conversation) {
  return check_aux(conversation, \(c) { return c.empty(); })
}
```

```
boolean check_aux(conversation, aux) {
  if (conversation.empty())
    return false;
  else if (conversation.getFirst().equals("STOP") ||
           conversation.getFirst().equals("RESP"))
    return aux(conversation.dropFirst());
  else if (conversation.getFirst().equals("REQ"))
    return check_aux(conversation.dropFirst(),
                     \(c) { return check_aux(c, aux); }
}
```

Here `\(c) { b }` denotes an anonymous function with a parameter `c` and body `b`.

Based on the program above we can formulate the following two business rules:
Main rule: *A conversation C is valid if it C is conditionally valid with a side rule "C is valid when empty"*
Meta rule: *A conversation C is conditionally valid in the presence of a side rule R if:*
(I) *C starts with* `STOP` *or* `RESP` *and the rest of C is valid according to R, or,*
(II) *C starts with* `REQ` *and the rest of C is conditionally valid with the* ***Meta rule*** *as a side rule*

The above rules seem quite complicated and non-comprehensible because of the presence of higher-order concepts. Nevertheless, they capture the complete semantics of the program above. We could derive simpler rules by simplifying the above to obtain, for example, the following ones
Termination: *A conversation consisting of* `STOP` *only is valid*
Unrequested response: *A conversation consisting of* `RESP` *followed by* `STOP` *only is invalid*
Unanswered request: *A conversation consisting of* `REQ` *followed by* `STOP` *only is invalid*
Request with response: *If we start with a valid conversation, then after inserting* `REQ` *in the beginning and* `RESP` *just before* `STOP`*, we still have a valid conversation.*

Naturally, the above rules do not comprehensively grasp the meaning of the program, but only some of its aspects. However, they have the advantage of being readable and understandable.

4 Conclusion

Most of the authors that published algorithms for extraction of business rules from source code escaped carefully to discuss the treatment of the loops of the program (see for example [Putrycz&Kark, 2007]) refusing in such way to extract rules encoded in the loops. As was shown above, using the refunctionalization approach the problem could be solved. Something more: the opposite process — defunctionalisation — could also be helpful. An interesting further task will be to try to interpret a set of business rules directly instead to transform them to requirements and to write programs based on these requirements. In such a case, defunctionalization of the declarative business rules could be used.

References

[Baxter&Hendrix, 2005] Baxter, I., St. Hendrix, *A Standards-Based Approach to Extracting Business Rules*, Semantic Designs Inc., 2005 (unofficial presentation).

[Bisbal et al., 1999] Bisbal, J., D. Lawless, B. Wu, J. Grimson. Legacy information systems: issues and directions. *Software* 16(5), 1999, 103–111.

[Danvy&Millikin, 2009] Danvy, O., K. Millikin. *Refunctionalization at work*. Sci. Comput. Program 74(8), pp. 534–549, 2009.

[Danvy&Nielsen, 2001] Danvy, O., L. R. Nielsen. *Defunctionalization at work*. Proceedings of the 3rd ACM SIGPLAN international conference on Principles and practice of declarative programming, pp. 162–174, 2001.

[Hay&Healy, 2000] Hay, D., K. A. Healy (eds.). *Defining Business Rules ∼ What Are They Really?*. GUIDE Business Rules Project Final Report, rev. 1.3., July, 2000.

[Manev et al., 2010] Manev, Kr., A. Zhelyazkov and St. Boychev, *Implementation of an Object-Oriented Test Data Generator*. Proc. of International Conference on e-Learning and the Knowledge Society – e-Learning'10, Riga, 2010, pp. 231–236.

[Manev et al., 2012] Manev, Kr., N. Maneva, H. Haralampiev. *Extracting Business Rules Through Static Analysis Of The Source Code*. In Mathematics and Education in Mathematics, Proc. of the 41-st Spring Conference of UBM, Borovetz, 2012, pp.247–253.

[Maneva&Manev, 2011] Maneva, N., Kr. Manev, *Extracting Business Rules — Hype Or Hope For Software Modernization*, International Journal „Information Theories and Applications", Vol. 18, Number 4, 2011, pp.390–395.

[McGee, 2005] McGee, J., *Legacy Systems: Why History Matters*, Enterprise Systems Journal, 2005, `http://esj.com/articles/2005/10/11/legacy-systems-why-history-matters.aspx`

[Putrycz&Kark, 2007] Putrycz, A.W., Kark, *Recovering Business Rules from Legacy Source Code for System Modernization*, LNCS 4824, pp. 107–118, 2007.

[SSA, 2012] *Smart Source Analyzer* (*SSA*), Musala Soft, Ltd.: `http://www.musala.com`

[Stineman, 2009] B. Stineman, *IBM WebSphere ILOG Business Rule Management Systems: The Case for Architects and Developers*. IBM Software Group, November 2009.

Krassimir Manev, Trifon Trifonov
Faculty of Mathematics and Informatics, Sofia University
5 James Bourchier Blvd., 1164 Sofia, Bulgaria
manev@fmi.uni-sofia.bg, triffon@fmi.uni-sofia.bg

Blueprint of an Experimental Cloud Framework[1]

Radko Zhelev

Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, 1113 Sofia
radko.jelev@gmail.com

Abstract. In this paper we propose distributed system architecture and the use cases of the experimental cloud framework Cl²OSD (Clustered Cloud-aware Open Service Directory). This architecture combines the benefits of peer-to-peer clusters with a two-level hierarchical topology, as an sub-optimal balance between scalability and efficiency for a resource management system inside the cloud. Cl²OSD is easy implementable and deployable system architecture. It is suitable for publishing, browsing and using system and application functionalities structured as services. This affects in hitting up the performance capabilities and the limits of scalability of our clusters. Our framework enables high performance nonreplica caches for arbitrary functional modules handling resources inside our topology.

Keywords: Cloud Systems, Resource Management, Distribution Topology, SOA.

1. Cloud Management Architectures

According [1] the system support of the most important features in cloud architecture are structured as few major system processes: monitoring of the resources and service calls; matchmaking decisions about the resource allocation or service discovery; and management decisions' actuation. Each of these system processes can be implemented as a centralized cluster-wide process or as a set of distributed processes which cooperate in a peer-to-peer (p2p) or other proper topology scheme.

The other important feature of the cloud management is the supported information level. One can either keep very precise information about the local load condition of the managed cluster nodes and about the incoming process of service calls; or alternatively the system information can be monitored less precisely and/or rarely. In the last case one shed the precision and correctness of the management decisions for the price of a smaller system overload. The set of system parameters proposed in [Haselt] can be used as pattern for studying and performance evaluation of most cloud architectures including the proposed in this paper.

Existing taxonomies of distributed systems [2, 3 and 4], initially classify

[1] This research is supported by the project ДДВУ 02-22/20.12.2010 of the National Science Fund.

the organization of resources into centralized and decentralized. Centralized management systems (as well as centralized clusters) have one single machine functioning as a central manager that handles the whole synchronization and work distribution at central place. Centralized management is simple and easy for realization, but suffers from lack of scalability due to the load capacity limitation of the central manager and cannot ensure reliability and high availability due to the single point of failure [5]. A taxonomy of Resource Management Systems [3] separates the decentralized organization of resources into flat, cells and hierarchical. In a flat organization all machines can directly communicate with each other without going through an intermediary. This organization we can refer as a peer-to-peer cluster. In a cell structure the machines within the cell communicate internally using flat organization, but there are designated machines acting as boundary elements that are responsible for all communication outside the cell. We can refer such structures as mediated clusters. Mediated clusters usually scale better than peer-to-peer clusters because of the minimized I/O synchronization. But for this scalability they pay with lower performance because of the intermediate redirections from boundary nodes to internal nodes and with lower availability since dropping of boundary elements causes loss of the whole cluster. In a hierarchical organization, machines in the same level can directly communicate with the machines directly above them or below them. Most current Grid systems use this organization since it has proven scalability. Of course as much the hierarchy levels are, as slower the performance becomes due to the increased number of redirections as requests travel forth and back. For a cluster level performance prediction we adopt the method described in [1]. During the development stage we use the simulation results of the models presented there as a

Further in this paper we present the system architecture, service infrastructure and the major use cases of the proposed experimental cloud framework.

2. Cl^{2}OSD System Architecture

2.1. General View

The Cl^{2}OSD infrastructure consists of a cluster of nodes with any-to-any connectivity and flat hierarchy – Fig. 1. The users of this system are allowed to publish services in the common list of services, or to launch any already published service. Each user is supported by a connection to a node of the cloud.

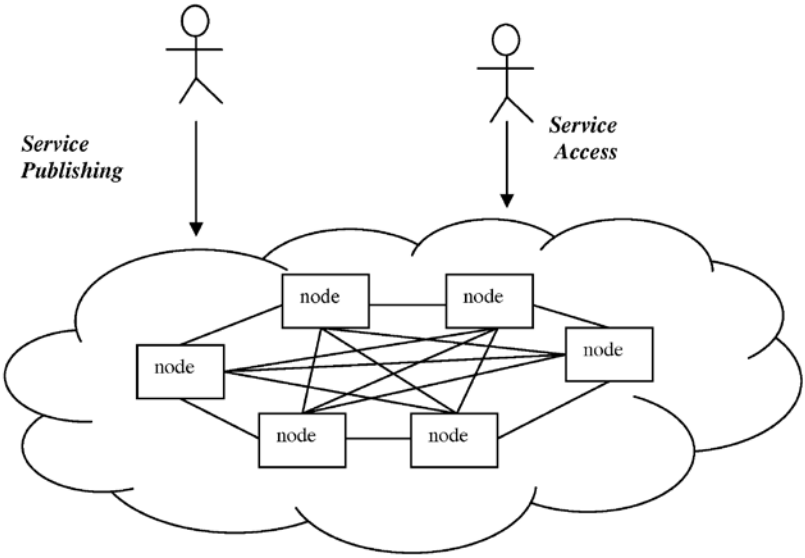


Fig. 1. Cl2OSD infrastructure of clustered nodes.

Each node is a JVM running Equinox OSGi framework. The nodes are organized in a cloud cluster by a set of system services that support cooperative parent to the cloud-unaware services, or available as regular application services building cloud-aware applications.

2.2. Services in Cl²OSD

In Cl²OSD we designate two general types of services regarding their role and purpose in the system. In one hand we have the so called System Services that build up and leverage the entire Cl²OSD infrastructure. In the other we have the End-User Services. They are published and shared by Cl²OSD Users and provide end-user functionality. To make a distinction – End-User Services are directly utilized by the Cl²OSD users, while System Services organize the underlying Cl²OSD middleware and are accessible programmatically.

Additionally, End-User Services are separated into Stand-alone and Cloud-aware Services. Cloud-aware Services run inside the Cl²OSD JVM, they are «aware» about the Cl²OSD environment and potentially may use any of the System Services to provide more extensive and complex distributed functionality.

An End-User Service in Cl²OSD can be generally:

- published – meaning that it is installed in the Cloud and available for usage by all Cl²OSD users;
- used by Cl²OSD users – meaning that a user can utilize the target Service functionality via the Cl²OSD Web Portal;

- uninstalled – removed from Cl²OSD, becoming no more available for usage.

Stand-alone services are simply executable binaries that can be normally started on Windows OS (Java programs or win executables) and provide some Web functionality for interaction with users.

In any case, stand-alone or cloud-aware Services need to provide a Web interface to their functionality, so that it would be possible to load the Service UI into the browsers of Cl²OSD Service Users.

2.3. Cl²OSD Architecture and System Services

Cl²OSD design strictly follows recent state-of-the-art SOA (Service Oriented Architecture) model for building applications. The whole Cl²OSD middleware is provided as loosely coupled units of functionality called System Services.

The system is supported by the following services. The Infrastructure Service supports dynamic information about the set of nodes running Cl²OSD in the cluster. This service recognizes two states of the nodes – up and down. The Communication Service supports peer-to-peer messaging in the cluster. The Information Service supports dynamic list of the published services and the set of services' attributes uding node[s] of current launching, initiating user[s], etc. Reservation & Scheduler Service.

Unlike the most SOA examples, Cl²OSD SOA architecture is built on the top of the so called OSGi Framework. OSGi provides several benefits that we consider crucial for Cl²OSD purposes:

- high dynamics in modularity – loading and lifecycle of modules;
- lightweight service registry;
- a specification standard enabling interoperability.

Software modules (in terms of OSGi Framework) are called Bundles and they can be dynamically installed, started, stopped, updated and uninstalled by the framework. Dynamic class loading and package exporting enables new bundles providing functionalities to be added without restarting of the Java VM. The Service Registry provides a cooperative environment for Bundles to register the so called OSGi Services. An OSGi Service represent a java object that implements arbitrary java interface defining the service functionality. Via the Service Registry, OSGi Services can be shared among the modules running inside the OSGi Framework (Fig 2).

There are several open source implementations of the OSGi Framework Specification and we have chosen to build Cl²OSD on the top of one of them, namely - the Apache Karaf OSGi Framework. There is nothing special in making this choice, and since OSGi is a specification standard, it should be easy and trivial to port Cl²OSD on any other OSGi Platform.

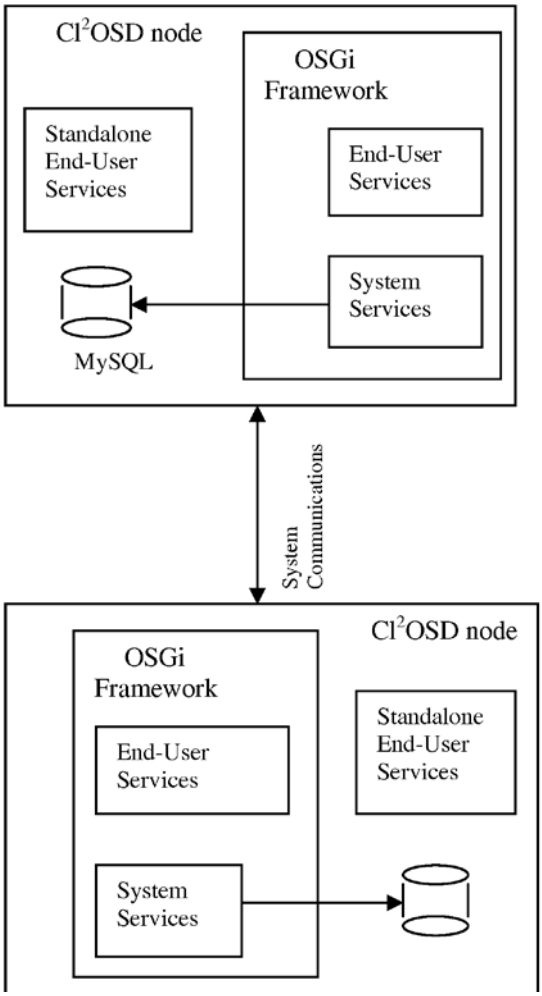


Fig. 2. Physical deployment over OSGi Framework.

2.4. System Services in Cl²OSD

All System Services are provided as OSGi Services registered in the OSGi Framework Platform on which Cl²OSD is running. Following is a list of the System Services building up the System. They internally access each other, cooperating together to handle different functional use-cases, but they can also be accessed by the End-User Services if desired (see 4.4.2 Using the System Services).

- Infrastructure Service – organizes the distributed infrastructure and provides information about the nodes participating in the System: identity and current state of any particular node;

- Communication Service – provides message-based communication between nodes;
- Storage Service – replica-enabled distributed storage (store here, find anywhere);
- Information Service – provides information about the End-User Services published in the System and User Accounts in Cl²OSD;
- Reservation & Scheduler Service – the load-balance decision module. Decides on which nodes End-User Services to be installed and which node(s) to handle particular service queries. Capable for handling the failover cases when physical failures occur;
- Front Service – two basic purposes: a) provides the Web UI of Cl²OSD – html-based user interaction for login, publishing and viewing of available Services, choosing of particular service to be used, etc.; b) provides http-bridge between the end-user service queries and the physical place where the end-user service instance is running inside the Cloud.
- HTTP Servlet Engine – this one comes with the base OSGi Framework underneath. Can be used by the cloud-aware Services (see 2.3.3 Cloud-aware and Stand-alone Services) to register their HTTP Servlets and provide interactive Web UIs of their own.

All system information together with Services' own data (handled via the Cl²OSD Storage Service) is kept within MySql Database servers running inside the Cl²OSD Nodes.

Each node is running one instance of the OSGi Karaf Framework with all Cl²OSD system Bundles installed, thus forming the Cl²OSD software platform. Each node is also running one instance of a MySql Database. Nodes communicate between each other exchanging messages only via the Communication Service and each node access db data only from the local MySql Server.

2.5. Virtual Cluster as a Service

When a Service is published into Cl²OSD, the System automatically decides on which nodes the service should be initially installed. The chosen nodes (3 in count by default) form the so called 'Virtual Cluster' of the Service. When user queries targeted to this Service come into Cl²OSD, they will be load-balanced in the scope of its Virtual Cluster, since those are the nodes where the Service binaries are installed and running. A number of 3 is chosen as an initial default Virtual Cluster size, since it is considered as best-balance between minimal resource occupations and high availability of the Service, i.e., as long as any of those 3 nodes is alive, the Service can be used by the Cl²OSD users. For future versions of Cl²OSD, automatic scale-up and scale-down of the Virtual Cluster size is planned, corresponding to the increasing or reducing of the load demand

on a particular Service as it becomes more or less popular to the users in the Cl²OSD Directory.

Cloud-aware services are developed and packed as OSGi Bundles. Publishing in Cl2osd means 'installing' of the OSGi Bundle into the base OSGi Framework where the System-bundles are also running (Fig. 2.). Thus, cloud-aware Services can in turn use the System Services (available in the OSGi Service Registry) for any specific purpose related to their functional needs. For instance they can exchange messages with the other Cl²OSD nodes, they can use the Cl²OSD Distributed Storage to save persistent data visible on all Service nodes (in their Virtual Cluster), they can use the OSGi HTTP Service as a Servlet engine/ container to export their Web Interfaces to Service users, and so on.

Stand-alone Services are simply java applications or any executables that can run on a Windows platform. These Services are not aware of the Cl²OSD environment and they make no difference if they are simply started within or outside the cloud. Such applications/services are provided to Cl²OSD packed as zip archives. Installing of a service means unzipping the archive on the respective Cl²OSD nodes and executing its startup script if any.

2.6. Services Support

In order to be handled and managed by Cl²OSD, all Services (stand-alone or could-aware) must contain ***metadata*** information describing how the Service should be treat. Metadata is provided as a simple property file, named '`cl2osd.props`', containing key-value pairs with predefined key property names. The '`cl2osd.props`' file must be located in the root directory of the Service archive, i.e., the bundle jar file for cloud-aware Services, or the generic archive for stand-alone ones.

In order to be accessed by Cl²OSD users, End-User Services must provide a Web Interface that can be loaded in the browsers of the Cl²OSD users. This is handled differently by the cloud-aware and stand-alone Services.

Cloud-aware service provide its Web Interface using the 'HTTP Service' servlet engine/container coming with the OSGi Framework (Note that HTTP Service is a standard OSGi Service specified by OSGi and available in any OSGi-compliant framework implementation including ours).

Stand-alone services can open a port and serve http queries like web servers, i.e., they can directly accept TCP-sockets and interact via the HTTP protocol; or can integrate (i.e. pack within their service archive) a whole Application Server like 'Glassfish' or a Servlet Container like 'Tomcat' for example, having their application installed within. They also can provide static HTML pages that will be automatically accessed by Cl²OSD in response to user queries addressed to that Service. In this way, Services can also supply their Web UIs as Java Applets that will be loaded into the user's browser.

3. Conclusions and Feature Work

The first task of our work is the definition of hybrid distributed architecture of an experimental multipurpose and highly adaptable cloud framework. It is a result of chasing after the best balance between scalability and performance efficiency for the cloud.

Current implementation of Cl^{2}OSD is accessed by its users via a dedicated Web Port UI Interface. Once a user loads the Cl^{2}OSD URL in its browser, a login page asks him to login into Cl2osd. After login, a list of published Services is shown, so the user can choose a Service to be used. On choosing a particular Service, Cl^{2}OSD loads the Service's Web UI into the user's browser, hiding from the user how and on which cluster Node its user session is being served by the real Service instance. The user queries are proxied by Cl^{2}OSD via an HTTP bridge implemented by the Cl^{2}OSD Front Service.

A matter of future work would be to provide experimental proof for the efficiency of the proposed architecture and the implementing framework. Parameters should be applied for a set of characteristics – the possible host states including lightly-loaded and heavily-loaded hosts, the probable load upon the system from administration side and from resource side, etc.

Another direction of research would be to modify the algorithm for heterogeneous clusters, i.e. composed of machines with different hardware capabilities that may handle different volumes of work. This is not a practical requirement for the Cloud, since Clouds are intentionally built and hence - dedicating a homogeneous cluster is not an issue. But our algorithm could still be modified to work for a heterogeneous cluster providing good utilization without monitoring the load of every machine.

References

1. Georgiev, V. Load Diffusion and Brownian Models for Cloud Balancing: between C-S and p2p. In Proceedings of ESM'2010, Hasselt, Belgium, 25–27. October, 2010. pp. 170 – 177.
2. Georgiev, V., J. Karvo. Numerical Modeling of Load-balanced Multicore Servers in a Cloud Cluster. Proceedings of 12th Middle Eastern Simulation and Modeling Multiconference MESM'2011, Amman –Jordan, November 14-16., 2011. pp. 48 – 52.
3. Rotithor, H. G., Taxonomy of dynamic task scheduling schemes in distributed computing systems, IEE Proceedings on Computer and Digital Techniques, 141, No. 1 (Jan. 1994), 1–10.
4. Yu, J. and R. Buyya, A Taxonomy of Workflow Management Systems for Grid Computing, Journal of Grid Computing, Volume 3, Numbers 3-4, Pages: 171-200, Springer Science+Business Media B.V., New York, USA, Sept. 2005.
5. Krauter, K., R. Buyya, M. Maheswaran, "A taxonomy and survey of grid resource management systems for distributed computing", Software: Practice and Experience. vol. 32, 2, 2002, pp. 135-164

Adaptive Integrated Business Management Systems

Milko Tipografov[1], Evgeniya Grigorova[1]

milko@pias-solutions.com, eva@pias-solutions.com

Abstract. The dynamics of the IT development market as well as the unexpected by many financial crisis have identified several 'must haves" when it comes to corporate purchasing of IT systems. This paper explores how modern business responds to threats of economic insecurity and globalization with the utilization of: integrated business management systems, new technologies (SaaS, SOA, CLOUD, N-tier Architecture), out-of-the-box process-oriented software solutions, "global best practices", and "agile" methodologies for deployment.

Keywords: SaaS, SOA, CLOUD Computing, ERP, CRM, BI

1. Introduction

Several years ago companies all over the world were realizing that business was becoming more and more mobile and that technology and business IT infrastructure were soon to be expected to deliver connectivity, speed, security and efficiency all on the go and at the lowest cost possible. Although some organizations were planning for the future most of them failed to anticipate, plan for, or react quickly to the banking crisis and the subsequent economic dislocation. This major event further increased competition and rates of change [2].

What is more evident than ever is that business and technology are now inextricably linked but technology seems to have become a major obstacle for business growth. Due to the way IT was traditionally sourced, a lot of companies found it inefficient, not synchronized and most importantly difficult and costly to change. A fundamental shift is already embraced by pioneering companies who understand that the IT organization has to change in order to become agile and ready for a rapid change. Becoming agile and adaptive requires implementation and integration of different methodologies [4] as well as many changes in variety of areas. The major factor here is the complexity.

In this article we review the main obstacles that have be overcome or changed for organisations to become agile and adaptive from point of view of IT strategy so it can align to the business goals.

2. Exploring the Complexity

The root of the problem is the complexity as we have mentioned above. With the current understanding for developing IT strategy we can find it in:

- Numerous, disconnected and not integrated systems and software;
- Management of huge software and hardware assets requires human and monetary resources. The heritage of business systems is massively complex and poorly integrated;
- Development of software from scratch carries risks and issues;
- Deployment and adaptation problems;
- Huge IT budgets, blurred CTO;
- How to measure the ROI
- The human factor.

The analysis of above mentioned topics shows that we need to change our understanding of building our IT strategies. Change but to what and how? The answer is *agility* and *adaptability*. The organization strategy should follow the business needs using proper approaches to react to the challenges quickly and in the frame of an optimized budget.

2.1 Agile Transformation

Agility and adaptability is of great importance for the success of every project. All IT organizations must change fundamentally to achieve the level of agility that is needed in our uncertain world [5].

Based on our research we can outline the main factors that should be changed for our IT strategy to become agile. So there are grouped by IT activities and processes - infrastructure, software licensing, business software, deployment methodologies, and design principles. For each activity we have summarized the main directions and principles that one modern IT strategy should fulfill.

- Using SAAS licensing model
- SOA based system processes
- CLOUD Infrastructure
- N-tier Software Architecture
- Out-of-the-box process-oriented & "global best practices"
- “Agile" methodologies for deployment
- Embracing the “Best World Practices” - Class Business Software CRM +ERP+BI+Mobile+e-Business to work together

Following these guidelines the organizations could reduce the risks and improve the quality of the complex projects.

2.2 SAAS – flexible licensing.

Based on the SAAS model for licensing we pay what we use. Upgrades for new software versions are included. The payments are spread during the period

– there is now need to invest huge amounts in the beginning – this gives us better budgeting and helps for easier migration to new versions.

Software as a service, sometimes referred to as "on-demand software", is a software delivery model in which software and associated data are centrally hosted on the cloud. SaaS is typically accessed by users using a thin client via a web browser.

SaaS has become a common delivery model for many business applications, including accounting, collaboration, customer relationship management (CRM), management information systems (MIS), enterprise resource planning (ERP), invoicing, human resource management (HRM), content management (CM) and service desk management. SaaS has been incorporated into the strategy of all leading enterprise software companies. One of the biggest selling points for these companies is the potential to reduce IT support costs by outsourcing hardware and software maintenance and support to the SaaS provider [8].

Benefits of using SAAS model are:

- Pay monthly
- Pay what is used
- Use latest versions
- Easy scalability

Software as a Service helps organizations to achieve their business goals at a cost typically less than paying for expensive hardware platforms, licensed applications, installation and maintenance.

2.3 System Processes - SOA Principles

We should design out software architecture and IT infrastructure according to the SOA model and principles [1]. Some of the key principles behind SOA are to construct technology in ways that 'de-couple' component IT functions (in other words reducing the interdependence between two or more parts of the system) and to allow these component functions to be leveraged as services more easily by other applications. The reason this is important for agility is that it allows IT components to make changes without the risk of affecting other parts of the systems landscape. Reusing interfaces and escaping "Spaghetti" effect are the main principles that should lead us.

In software engineering, a service-oriented architecture (SOA) is a set of principles and methodologies for designing and developing software in the form of interoperable services. These services are well-defined business functionalities that are built as software components (discrete pieces of code and/or data structures) that can be reused for different purposes. SOA design principles are used during the phases of systems development and integration.

SOA generally provides a way for consumers of services, such as web-based

applications, to be aware of the available SOA-based services. For example, several disparate departments within a company may develop and deploy SOA services in different implementation languages; their respective clients will benefit from a well-defined interface to access them. XML is often used for interfacing with SOA services, though this is not required.

SOA defines how to integrate widely disparate applications for a web-based environment and uses multiple implementation platforms. Rather than defining an API (Application Programming Interface), SOA defines the interface in terms of protocols and functionality. [9].

The key benefits of using SOA principles are:

- Avoid the 'Spaghetti' problem
- De-couple processes
- Re-use interfaces and functionality

These are the reasons for many organizations to choose this advanced architectural approach.

2.4 CLOUD Infrastructure

Cloud computing is IT-as-a-Service. Instead of building your own IT infrastructure to host databases or software, a third party hosts them in its large server farms [6]. Your company has access to its data and software over the Internet.

Key benefits of using CLOUD Infrastructure are:

- Cheap:sharing complex infrastructure and using SaaS licensing is cost-efficient and you pay only for what you actually use.
- Quick: The most basic cloud services work out of the box; for more complex software and database solutions, cloud computing allows you to skip the hardware procurement and capital expenditure phase – it is perfect for start-ups.
- Up-to-date: Software offerings are constantly updated by adding new features as they become available.
- Scaleable: If your business is growing fast or has seasonal spikes, you can go large quickly because cloud systems are built to cope with sharp increases in workload.
- Mobile:Cloud services are designed to be used from a distance, so if you have a mobile workforce, your staff will have access to most of your systems on the go.
- Green: electricity consumption is optimized in data centers
- Secure: you get back up and recovery strategy on the start

As a whole, the CLOUD computing has a big future and more and more

companies will turn to it. The CLOUD services and used applications are to be discovered and explores by the modern business. It gives security, cheap scalability in all aspects of our IT, ability quickly to react and optimization of our IT budgets.

2.5 "N-tier versus Client-Server"

The main principle behind N-tier architecture is that the presentation, the application processing, and the data management are logically separate processes. N-tier application structure implies the client/server program model. Where there are more than three distribution levels or tiers involved, the additional tiers in the application are usually associated with the business logic tier.

The main advantage of n-tier architecture is that we can manage business process levels better and independently from interfaces and data storage levels – in one word it gives us flexibility. In addition to the advantages of distributing programming and data throughout a network, n-tier applications have the advantages that any one tier can run on an appropriate processor (CPU, machine, workstation) or operating system platform and can be updated independently of the other tiers.

2.6 One integrated solutions Business Software should work together. Out-of-the-box process-oriented and «global best practices".

For the modern reality it is not enough to implement CRM, ERP and BI solutions according the "Best World Practices" but to make them work together and open your company to the world and technologies that are to come - embrace new channels of communication and modern business strategies. So choosing a business software solution, vendor and underling technology stack together with scalability (in terms of new functionalities and business processes availability) and time for delivery are very important factors for your business software to last the future businesses threats and technology challenges.

2.7 Waterfall versus Iterative approach for deployment

Many projects based on the "Waterfall" model have failed because what was delivered is not what was expected from the process owners. A 'Waterfall' project would usually takes longer time and would cost more than initially planned. In projects and releases, the 'Waterfall' approach can give a false sense of security and control. Progress is measured using a number of 'input' and 'process' deliverables but, in truth, these are not robust indicators of delivery progress or quality.

The right answer is again "Agile" – early involvement of the process owners

and decision makers in the process of development and delivery is the most important factor for success. Here the key factors and techniques are the spiral model of development, horizontal delivery, prototyping and early feedback. They will give all the stakeholders real understanding and feeling of the expected system. The big gap between all the parties' understanding will be avoided and minimized.

3. Conclusions and Future Work

In conclusion we can say that the IT strategy should follow and align to business needs and should be ready to face the business threats and future technology challenges. It should be agile and adaptive, innovative, budget optimized, predictive and easy to manage. It should not be an obstacle for business growth and change but rather to give the business urge for change and stability.

We will continue to review and explore all the aspects of IT - activities, methodologies, good practices and technologies that should be changed or adopted by modern business and the way these factors influence the business goals and strategy.

References

[1] Димитров В., Ориентирана към услуги архитектура в рвализацията на MSIS 2006, Сборник доклади на международната научна конференция УНИТЕХ'09, 20-21.11.2009, Габрово, България, стр. III-372 – III-378

[2] Organisational agility: How business can survive and thrive in turbulent times, http://www.emc.com/collateral/leadership/organisational-agility-230309.pdf

[3] Kaloyanova K., Some aspects of implementing ITIL, Proceedings of the 6-th Annual International Conference on Computer Science and Educaton in Computer Science, 26-29 June 2010, Fulda-Munich, Germany, pp 48-53

[4] Napoli, J.P, K. Kaloyanova. An Integrated Approach for RUP, EA, SOA and BPM Implementation, Proceedings of the 12th International Conference on Computer Systems and Technologies, Vienna, Austria, 16-17 June 2011, ACM DL, ISBN: 978-1-4503-0917-2, pp.63-68

[5] IBM : Agile Transformation – rethinking IT strategy in an uncertain world, http://www-304.ibm.com/easyaccess/fileserve?contentid=208473

[6] IBM : Perspective on Cloud Computing, ftp://ftp.software.ibm.com/software/tivoli/brochures/IBM_Perspective_on_Cloud_Computing.pdf

[7] IBM : The evolving role of the CIO, http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-global-it-risk-study.html

[8] Software as a service, http://en.wikipedia.org/wiki/Software_as_a_service

[9] Service-oriented architecture, http://en.wikipedia.org/wiki/Service-oriented_architecture

Algorithm for Simulating Timestamp Ordering in Distributed Databases[1]

Svetlana Vasileva[1]

[1] Shumen University „Bishop Konstantin Preslavski“, College – Dobrich, Dobrotitsa 12, Dobrich, 9302, Bulgaria
svetlanaeli@abv.bg

Abstract. One of the major problems in Distributed database management systems is the concurrency control. The paper considers an approach for modeling management of the transactions parallelism by the timestamps method. As deadlocks originate in concurrency control algorithms, based on two-phase locking protocols, in order to avoid this problem, we offer timestamp ordering (TO) service of the transactions. The TO method in distributed databases is still not investigated enough. However, the use of timestamps makes the algorithms for transaction management more complex due to the restarting of the transactions from service and the additional waiting for processing. This is one of the reasons we consider a modeling algorithm of TO in greater details. The results from executing modeling algorithm by means of the simulations GPSS World environment are shown.

Keywords: distributed databases, distributed transactions, concurrency control, timestamp ordering, deadlocks.

1 Introduction

The distributed database (DDB) is a distributed system, which is an aggregation of logically connected local databases, geographically distributed and unified by a computer network. Some of the most important functions of the systems for distributed database management are: synchronization of the application processes that function in the distributed system and supply of the fault tolerance of the system. The application processes in the DDB system have to be managed in such a way so that the system remains whole even after emergency. There are developed different methods for concurrency control in order to provide wholeness of the information. The basic methods are: two-phase locking (2PL), timestamp ordering (TO) and optimistic strategies (validation check up) [1], [2] and others. According to a number of authours the best of the three methods with respect to the response time index in high workload in DDB system is the two-

[1] This paper supported by Project N РД-05-138 of Shumen University “Bishop Konstantin Preslavski” whose topic is Models and applications of the information technology and systems in education

phase locking (2PL) protocol in its variants: Centralized 2PL, Primary copy 2PL and Distributed 2PL. But the concurrency control by the pessimistic protocols has one main problem – the possibility of the transactions to be involved in the deadlocks. Therefore together with the basic 2PL protocols for DDB there are used algorithms for deadlock avoiding or deadlock solving.

The paper is dedicated to the investigations [5], [6], [7] of concurrency control protocols for DDB system. There is presented an algorithm modeling timestamp ordering in distributed database management system (DDBMS). The purpose is the different protocols for concurrency control and their variations to be investigated in details and to be compared in identical initial conditions in the DDB system and different regimes of workload.

2 Timestamp ordering method in distributed database systems

In the centralized database system the task of timestamp (TS) protocols is the global alignment of transactions so that the older transactions (which have smaller TS) in case of conflict to receive priority [1], [2], [8], [9] and others. The general approach in the DDBMS is concatenation of local timestamp with the unique identifier of the node (*<local TS>*, *<node identifier>*) [1], [9]. The node identifier value has a smaller weight coefficient which guarantees the order of the events in accordance with the moment of their appearance.

The serving of global transaction in the distributed timestamp ordering (DTO) modeling algorithm is performed according to the algorithm of timestamps, shown in fig. 1. The schema in fig. 1 demonstrates TO algorithm in a summary, described in [1], [9] and by other authours. The algorithm uses Tomas rules according to which:

- To each transaction *T* is assigned timestamp, denoting the time of its coming into the system and the number of the site-generator. When a transaction read/write data element, it records its TS in it;
- If a transaction ***T*** wants to update data element ***x***:
  If TS(***T***) < readTS(***x***), then restart(***T***);
  If TS(**T**) < writeTS(***x***), then ignore(***T***);
  If TS(***T***) > writeTS(***x***), then execute(***T***);
- If a transaction wants to read data element ***x***:
  If TS(***T***) < writeTS(***x***), then restart (***T***);
  If TS(***T***) > writeTS(***x***), then execute(***T***).

3 Basic Elements of the GPSS Simulation Model

The suggested Timestamp ordering algorithm is realized with the help of the simulation environment GPSS World Personal Version. The presented simulation model uses generated streams of transactions which simulate global transactions

in DDB systems. Simulation models for such initial conditions are presented in [5], [6] and [7]. Here we will emphasize the use of timestamps for distributed transaction concurrency control. The transactions are all in parallel streams and their intensity λ is given in tr per sec (number of transactions per second).

The structural scheme of a modeling algorithm for distributed transactions management in timestamp ordering algorithm in DDB is shown in fig. 1.

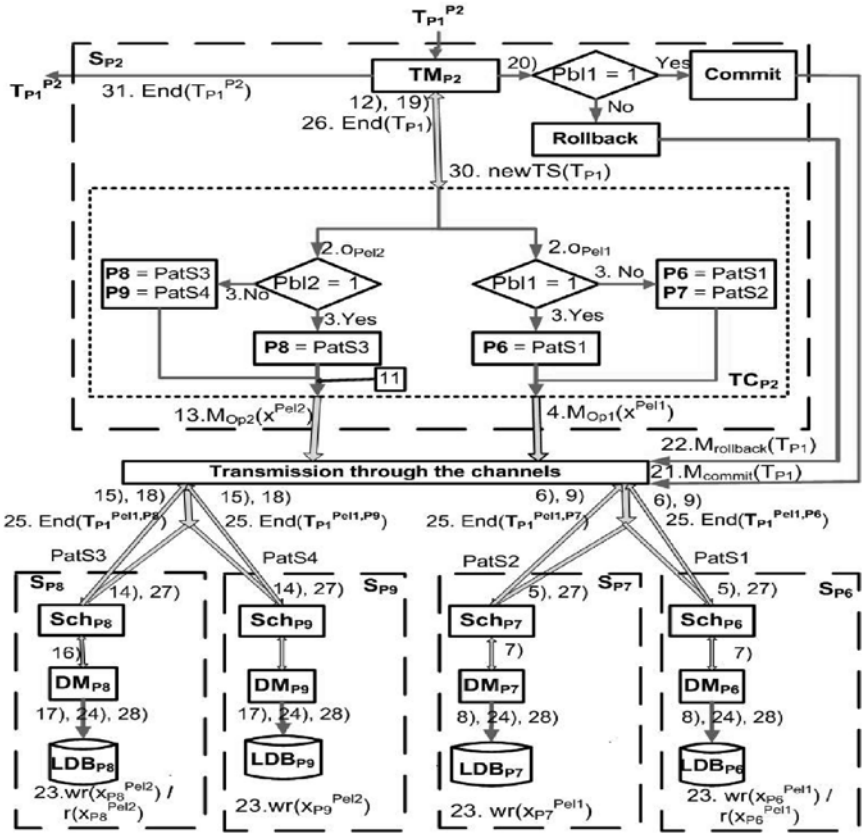


Fig. 1. A frame scheme of global transaction execution by the modeling algorithm of timestamp ordering

3.1 Parameters of the GPSS transactions

The parameters of generated GPSS transactions are the same as in [5], [6] and [7] except for the parameter P$Vr:

P1 – Number of transaction. The value is a sum of System Numeric Attribute MP2 (The subtraction between the relative model time and the content of the second parameter of GPSS transaction) and the number of the site;

P1 – number of the generating transaction;

P2 – number of the generating transaction site;

Pel1 / Pel2– number of the first / second processed data element by the transaction (El1) / (El2);

Pbl1 / Pbl2 – type of the operation over the element *El1* / *El2*: 1 (r) – if read (*El1*) / (*El2*); 2 (w) – if write (*El1*) / (*El2*);

P5 – phase of the transactions processing: it takes the value of 0 in the transaction coming in the model and after the end of the operation read/write it takes the value of 1. In the DTO model P5=2, if Ignore(*T*); P5=3, if Rollback(*T*);

P6 / P7 – number of the site where the first / second copy of the first data element *El1* is stored;

P8 / P9 – number of the site where the first / second copy of the second data element *El2* is stored;

P11 – number of the user's list where the corresponding sub-transaction waits for the release of the copy data element.

P$Vr – parameter that is used in making the decision about commit/rollback of transaction in DTO model: P$Vr=0, if the transaction has not requested the element yet; P$Vr=1, if T continues execution; P$Vr=2, if Rollback(*T*).

3.2 Basic operations

The basic steps in the synthesized DTO modeling algorithm (in fig. 1) are:

- Operation **1**: Coming of *T* into the system and processing of transaction manager;
- Operation **2**: Preparation for splitting *T* in transaction coordinator;
- Operation **3**: If the operation is for writing, then for the two copies of *El1* and *El2* T it must be split into sub-transactions;
- Operation **4**: Sending of the request for read/write of element *El1*;
- Operation **5**: Processing in schedulers of first and second replica of *El1*;
- Operation **6**: TS(*T*) < readTS(*El1*), then sending a restart message (*T*);
- Operation **7**: If TS(*T*) > readTS(*El1*), then *El1* is taken;
- Operation **8**: Temporary appropriation TS(*T*) of the copy of *El1*;
- Operation **9**: Sending a confirmation message of executing read/ write(*El1*);
- Operation **10**: If *El1* is not free, *T* queues before it;
- Operation **11**: Confirmation about continuation of the *T* execution;
- Operation **12**: Restart(*T*) – sending messages from schedulers to the transaction manager;
- Operation **13**: sending a request for read/write of element *El2*;
- Operation **14**: Processing in schedulers of first and second replica of *El2*;
- Operation **15**: TS(*T*) < readTS(*El2*), then sending a restart(*T*) message;

- Operation **16**: If TS(*T*) > readTS(*El2*), then *El2* is taken;
- Operation **17**: Temporary appropriation TS(*T*) of the copy of *El2*;
- Operation **18**: Sending confirmation message about execution of read/ write(*El2*);
- Operation **19**: Uniting and voting of sub-transactions of *T* for commit(*T*) or rollback(*T*);
- Operation **20**: Making a decision about commit(*T*) or rollback(*T*);
- Operation **21**: Sending a message to commit(*T*);
- Operation **22**: Sending a message to rollback(*T*);
- Operation **23**: Fixing the results from the execution of *T* in local databases;
- Operation **24**: Releasing of replicas of *El1* and *El2* from committed *T*;
- Operation **25**: End of sub-transactions that processed *El1* and *El2*;
- Operation **26**: End of *T*;
- Operation **27**: Abort(*subT*) over *El1* and Abort(*subT*) over *El2*;
- Operation **28**: Releasing of replicas of *El1* and *El2* from aborted *T*;
- Operation **29**: Confirm rollback(*subT*) over *El1* and Confirm rollback(*subT*) over *El2*;
- Operation **30**: Start(*T*) with new timestamp;
- Operation **31**: *T* is served and leaves the system.

The operation concerning the ignoring of *T* are not shown because they are modeled as execution without anything being recorded in the local databases.

The transfer through the network to the sites-executors, where are the schedulers and data managers (DM) is simulated with retention.

4 Simulation

Some fragments of a program GPSS World code, specific for the distributed transactions modeling and their service in distributed timestamp ordering are given here. We view a model of DTO synthesized similar to the modeling algorithms in [5], [6] and [7].

Example of a Computer Program from Vasileva S. (2012) GPSS World model of Distributed timestamp ordering

```
  …
DistrS1 FUNCTION V$SiteRepl1,D6  ;Replication of the data
1,2/2,6/3,1/4,5/5,3/6,4          ;First node for the El1
…  ;Replication of the data - Second node for the El1
...
BrEl FUNCTION RN4,D2    ;Number of the elements accessed
.30,1/1.,2              ;transaction
; Description of the used matrices
GBDA1 MATRIX ,50,5;modeling Local DB for first copies of El1 and El2
```

```
GBDA2 MATRIX ,50,5 ;modeling LDB for second copies of El1 and El2
…
INITIAL X$BROITR1,0  ;Counter of short transactions with length
INITIAL X$BROITR2,0  ;Counter of T with length 2 elements
INITIAL X$BROITR,0   ;Counter of generated transactions
INITIAL X$ZAVTR,0    ;Counter of commited transactions
;INITIAL X$VOT1, X$VOT2, X$VOT, X$VOT12 - the variables serving in
;the taking of decision for continuation, ignoring or rollback of T
INITIAL X$RESTRT,0  ;Counter of restarted transactions
; Description of the storages
; Description of the tables for statistic
*** Segment for transaction generation
        GENERATE 60,FN$XPDIS      ;Generation of
Potok1  ASSIGN 2,1                ;GPSS transactions with the
        ASSIGN 1,(MP2+FN$NomSait)  ;given incoming
        TRANSFER  ,BEGI            ;interval (60 ms) operation 1
BEGI  GATE FV FN$TraMan        ;Processing in
      SEIZE FN$TraMan          ;transaction manager
   FUNAVAIL FN$TraMan
   ASSIGN Nel,FN$BrEl     ;Giving the number of the elements
   ASSIGN El1,V$ElemN1    ;Giving the number of the El1
   ASSIGN 3,FN$LockTip1   ;Operstion type of El1
   ASSIGN 5,0           ;Flag of transaction committing
...
     SAVEVALUE BROITR+,1   ;Count generated transaction
…
Obrab  QUEUE FN$TransCor
   GATE FV FN$TransCor    ; Proccessing in Transaction
   SEIZE FN$TransCor      ;coordinators
…;Preparation for (not)splitting of transaction to subT
…;Defining values for P6, and for P7, and for P8, and P9
Pered2  RELEASE FN$TransCor
   FAVAIL FN$TransCor
Rolb12  SPLIT 1,Pat01,REPLI   ;Operation 4
        TEST NE P$Bl1,1,Per2  ;Splitting of the first subT
        SPLIT 1,Pat2,REPLI    ;operation 4
************* Operation 11 (confirm El1) or rollback T(El1)
Per2    GATHER 2      ;waiting for subT(El1) and T
        ADVANCE 1
        GATHER 2
        GATE SE GLAS
        GATE SV GLAS
        ADVANCE 1
        TEST NE X$Vot12,0,Cak12
        SAVEVALUE Vot12,0
Cak12   GATHER 2
        GATE SNF GLAS
        GATE SV GLAS
        ENTER GLAS
        SAVEVALUE Vot12+,P$Vr      ;Vot12 <2, if T will execute
        TEST E S$GLAS,2,
```

```
        SUNAVAIL GLAS
        ADVANCE 1
        GATHER 2
        LEAVE GLAS
        TEST NE X$Vot12,2,AborT01  ;If Vot12!=2, then T continues
        TEST E S$GLAS,0,Pat1ob
        SAVEVALUE Vot12,0
        SAVAIL GLAS
Pat1ob   TEST NE P$Nel,1,PrKrai
        GATHER 2
        ASSEMBLE 2
Per24   SPLIT 1,Pat3,REPLI         ;operation 13
        TEST NE P$Bl2,1,KraiT2
        SPLIT 1,Pat4,REPLI         ;operation 13
****Check abort or commit T after processing of the requests for
KraiT2  GATHER 2           ; read/write (El2)
        ADVANCE 1          ;waiting for subT(El2) and T
        GATHER 2
        GATE SE GLASUVANE
...
        ENTER GLASUVANE
        SAVEVALUE Vot+,P$Vr     ;If Vot12 <2, T will execute
...
        SAVAIL GLASUVANE
PatOb   GATHER 2
        ASSEMBLE 2
        TRANSFER ,PrKrai
AborTE  TEST E S$GLASUVANE,0,ObedAbort
...
ObedAbort  GATHER 2
          ASSEMBLE 2
        ASSIGN 5,3        ;If T has to be restarted, then P5=3
        ASSIGN Vr,2       ;and P$Vr=2
        SAVEVALUE RESTRT+,1        ;Counting of restarted T
******rollback subT(El1)
     SPLIT 1,Pat1,REPLI ;operation 22 - message_rollback to Sch_P6
        TEST NE P$Bl1,1,KrRest
     SPLIT 1,Pat2,REPLI ;operation 22 - message_rollback to Sch_P7
KrRest  TRANSFER ,Rolb1    ;T will be restarted
PrKrai  TRANSFER ,Krai     ;T will be committed
**********************
AborT01   TEST E S$GLAS,0,Pat1abo   ;Rollback(T) after
        SAVEVALUE Vot12,0           ;processing of El1 copies
        SAVAIL GLAS
Pat1abo GATHER 2
         ASSEMBLE 2          ; T will restart
        SAVEVALUE RESTRT+,1
Rolb1   ASSIGN 1,(MP2+FN$NomSait)  ;Taking of the new time stamp
        TRANSFER ,Rolb12           ;operation 27 restart(T(El1)
Pat1   ADVANCE MX$RAZST(P2,P6),MX$RAZDEV(P2,P6) ;Transfer to the
LockR1  ASSIGN NoCop,1                          ;executor-site(P6)
```

```
     QUEUE P6 ; Waiting in front of Scheduler of first copy of El1
        SEIZE P6   ;Processing start in scheduler P6 - operation 5
        DEPART P6
        ADVANCE 3,1
        RELEASE P6
ProvFiks  TEST E P5,0,Fiksira1
        ADVANCE 2,1
        TEST E CH*El1,0,Chakane1
   TEST E MX$GBDA1(P$El1,2),0,Chakane1 ;If El1 is free, T takes it
Zaema1  TEST NE P$Bl1,1,ObrChet1
     TEST G P1,MX$GBDA1(P$El1,3),Restrt11  ;If TS(T)>writeTS(El1)
     TEST G P1,MX$GBDA1(P$El1,4),Ignor1   ;If TS(T)>read(El1)
  MSAVEVALUE GBDA1,P$El1,2,P$Bl1 ;record in the LDB, that T will
        TRANSFER ,PatPot1                  ;write first copy El1,
ObrChet1 TEST G P1,MX$GBDA1(P$El1,4),Restrt11 ;If TS(T)>readTS(El1)
      MSAVEVALUE GBDA1,P$El1,2,P$Bl1  ;record that T will read El1
PatPot1 SPLIT 1,Pat1pot,REPLI        ;SubT is directed to DM_P6
        TRANSFER ,PatLB1    ;
Ignor1  ASSIGN 5,2         ;Ignore(subT), P5=2
        TEST NE P$Nel,1,Zakr1
        SPLIT 1,Pat1Pot,REPLI
        TRANSFER ,Zakr1
Pat1Pot ADVANCE MX$RAZST(P6,P2),MX$RAZDEV(P6,P2)
        TRANSFER ,PotRest11
Restrt11 ASSIGN 5,3     ;Abort(subT), P5=3
       ADVANCE MX$RAZST(P6,P2),MX$RAZDEV(P6,P2) ;towards making a
        TRANSFER ,PotRest11     ;decision about abort/continue(T)
Chakane1 LINK P$El1,FIFO
Fiksira1 TEST NE P5,3,Osv1   ;record in the LDB_P6 to unlock(El1)
         TEST E P$Bl1,1,RTSX1          ;fixing in the LDB
         MSAVEVALUE GBDA1,P$El1,3,P1   ;the results of subT over
         TRANSFER ,Osv1                ;first copy of El1
RTSX1    MSAVEVALUE GBDA1,P$El1,4,P1
Osv1     MSAVEVALUE GBDA1,P$El1,5,P2  ;operation 28
         MSAVEVALUE GBDA1,P$El1,2,0
         QUEUE QSEGMRW
         ENTER SEGMRW
         DEPART QSEGMRW
         TEST G CH*El1,0,Fik1
         UNLINK P$El1,Provch01,1
Fik1      LEAVE SEGMRW
         ADVANCE 3,1
         TRANSFER  ,Krai1
Provch01   TEST E P$Bl1,1,Izliza1
         UNLINK E P$El1,Provch01,1,Bl1,1,Izliza1
Izliza1  TEST E P$NoCop,1,Zaema3
         TRANSFER ,Zaema1
Krai1    ADVANCE MX$RAZST(P6,P2),MX$RAZDEV(P6,P2)
         TRANSFER ,Krai01
*****************************
Pat2  ...  ; Transfer to the executor-site(P7)
```

```
…
Fiks2   …;record in the LDB_P7 to unlock(El1)
…
PotRest11 TEST E P$Bl1,1,PotRest1   ;Check of Restart or
          TEST NE P5,3,DrT11        ;continue execution of T
          ASSIGN Vr,1        ;P$Vr=1 if P5!=3
      TRANSFER ,Per2  ;to voting for abort/execution subT(El1)
DrT11     ASSIGN Vr,2
          TRANSFER ,Per2
PotRest1  GATHER 2
          ADVANCE 2
          GATE SE GLASUVANE1
          GATE SV GLASUVANE1
          ADVANCE 1
          TEST NE X$Vot1,0,Cak1
          SAVEVALUE Vot1,0
Cak1      GATHER 2
          GATE SNF GLASUVANE1
          GATE SV GLASUVANE1
          ENTER GLASUVANE1
          TEST NE P5,3,DrugTr1
          SAVEVALUE Vot1+,1     ;Vot1=2, if T will execute
          TRANSFER ,Vote1
DrugTr1   SAVEVALUE Vot1-,1      ;Vot1 <2, if T will restart
Vote1     TEST E S$GLASUVANE1,2,
          SUNAVAIL GLASUVANE1
Reshen1   ADVANCE 1
          GATHER 2
          LEAVE GLASUVANE1
          TEST E X$Vot1,2,AbortT1
          TEST E S$GLASUVANE1,0,Prod1
          SAVEVALUE Vot1,0
          SAVAIL GLASUVANE1
Prod1     ASSEMBLE 2
          ASSIGN Vr,1     ;T continues
          TRANSFER ,Per2
AbortT1   TEST E S$GLASUVANE1,0,Prad1
          SAVEVALUE Vot1,0
          SAVAIL GLASUVANE1
Prad1  ASSEMBLE 2
       ASSIGN Vr,2      ; T will restart and P$Vr=2
       TRANSFER ,Per2  ; To voting for abort/execution subT(El1)
*******************************************
Pat3  ADVANCE MX$RAZST(P2,P8),MX$RAZDEV(P2,P8)  ; Processing as
…                                              ;in the Pat1 segment
Pat4 ADVANCE MX$RAZST(P2,P9,MX$RAZDEV(P2,P9  ; Processing as
…                                              ;in the Pat2 segment
****************************************
PatLB1  QUEUE FN$Opash1        ; operations 8, 23, 24 in LDB_P6
        ENTER FN$Opash1
        DEPART FN$Opash1
```

```
        ADVANCE 3,1
        TEST E P5,0,KrAb1
        TEST NE P$Bl1,1,FiksRW11
        MSAVEVALUE GBDA1+,P$El1,1,1   ;operation 8, 23, 24
        TRANSFER ,FiksRW11
KrAb1   TEST NE P$Bl1,1,Abor11
        TEST E P5,3,Abor11
        MSAVEVALUE GBDA1-,P$El1,1,1  ;restoration of El1 value
        TRANSFER ,Abor11
FiksRW11 ASSIGN 5,1
Abor11   LEAVE FN$Opash1  ;
Zakr1  TRANSFER ,ProvFiks
****************************
PatLB2  ... ;operations 8, 23 and 24 in LDB_P7
PatLB3  …   ;operations 17 and 24 in LDB_P8
PatLB4  …   ;operations 17 and 24 in LDB_P9
********************8********
Krai01 TEST NE P$Bl1,1,PKrai
          GATHER 2
          ASSEMBLE 2
          TRANSFER ,PKrai
Krai02 TEST NE P$Bl2,1,PKrai
          GATHER 2
          ASSEMBLE 2
    PKrai  TEST NE P$Nel,1,Krai
           ASSEMBLE 2
    Krai   GATHER 2        ;Merging of successfully
           ASSEMBLE 2      ;finished their work(sub)transactions
           TEST NE P$Nel,1,Saber1
           SAVEVALUE ZAVTR2+,1   ; Count committed transactions
           TRANSFER ,Napus       ;accessed two elements
    Saber1  SAVEVALUE ZAVTR1+,1  ;accessed only one element
    Napus  SAVEVALUE ZAVTR+,1  ;Count all committed transactions
            TABULATE TablZav
            DEPART TOTALTIM
            TABULATE DaTable
           TERMINATE 0
    *********************
         GENERATE 28800
         TERMINATE 1
```

5 Simulation Results

The parameters and indexes of the simulations of the considered model are as follows: *NumTr* – general number of the generated transactions for the time of incoming modeling; *FixTr* – general number of the completed (committed) transactions for the same period; *X=FixTr/Tn* – throughput of the queuing system; *Tn* – time interval in which the system is being watched; *Ps=FixTr/NumTr* – probability for transaction service. The results are received in 6 streams of

concurrent transactions with different intensity as in [5], [6] and [7]. The copies of the data elements are distributed evenly and random by 6 sites in the system.

The results of our simulations of the model for equal intensity of 6 input flows for 2 element copies are summarized and presented graphically in fig. 2 - fig. 3. The similar summarized results for different intensity of streams are shown in fig.4 – fig 5. The diagrams show the results of the conducted simulations of the presented GPSS model of primary copy 2PL with TS in DDB [6] for different input intervals of transaction coming.

Fig. 2 shows the results from throughput for simulation of distributed timestamp ordering algorithm in same intensities of the incoming streams of global transactions. It is seen that the throughput has the same behavior of increment in transition mode as the system load increases. For static mode the throughput is constant and it has max value.

The graphics in fig. 3 show the results of probability service for simulation DTO algorithm in the same intensities of the incoming streams of global transactions. The graphics in fig. 2 and fig. 3 represent the processing in the model for different loads: min, average and max. The graphics in fig. 3 show, that the probability service has the same behavior when the monitoring time is increased and the intensity of input streams accepts three different values. Moreover, the probability service reaches the max possible value when the static mode is used.

Fig. 4 shows the results of throughput for simulation of DTO algorithm in different intensity distribution of the incoming streams with one and the same summary intensity 60,87 tr/s. The diagram on fig. 4 confirms the results given in the fig. 2. The intensity of the output stream with processed transactions retains its behavior when the intensity of summary input stream is constant. It is seen a sharp increment in the beginning of monitoring time interval and a constant max value for static mode.

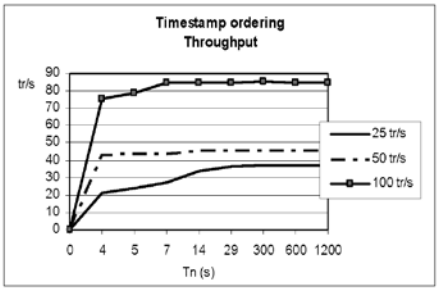


Fig.2. Throughput of the model in one and the same intensities of the incoming streams

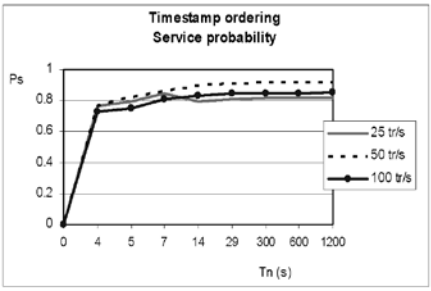


Fig.3. Probability service of the model in one and the same intensities of the incoming streams

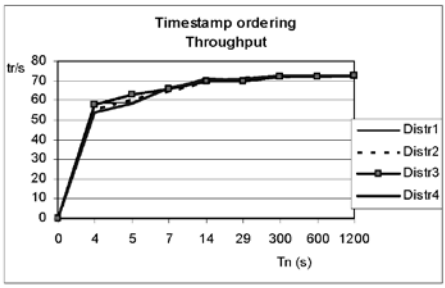


Fig. 4. Throughput of the model in different intensities of the incoming streams with one and the same summary intensity 0,6 tr/s

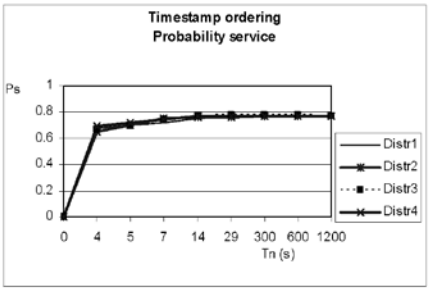


Fig. 5. Probability service of the model in different intensities of the incoming streams with one and the same summary intensity 0,6 tr/s

The graphics in fig. 5 show the results of probability service of distributed transactions for simulation of DTO for different distribution of the intensity for the six input streams with generated transactions from fig. 4. It's clear that the probability service depends only from the summary intensity of the input stream when the transition mode is running (fig. 5).

Fig. 6 shows the diagram of frequency distribution of response time (in case of: summary input intensity 100 tr/s, modeling time 28800 model units, i.e. just before stationary regime). The diagram in fig. 6 corresponds to the stereotyped graphic of response time, shown in [4, p.74].

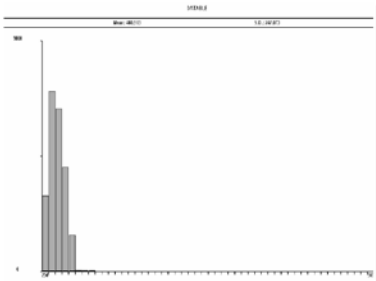


Fig.6. Frequency distribution of transaction RT in DTO model for input intensity 100 tr/s.

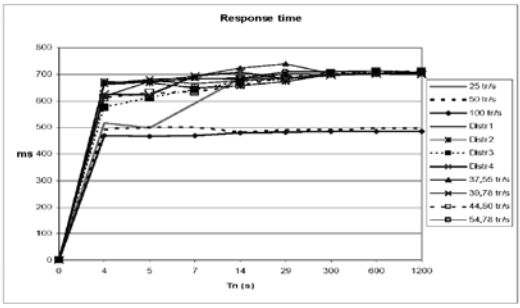


Fig.7. Response time (RT) of the model in the different intensities of the incoming streams

The diagram in fig. 7 summarizes in a comparative way the results for response time in different intensities of the incoming stream of global transactions. The results for response time show instability – when the intensity of the incoming stream of transactions grows, the response time also grows.

6 Conclusions

It is suggested a structural scheme of modeling algorithm to control the transactions by timestamp ordering protocol in distributed management database systems.

A program code for GPSS World is developed which can be used to model the processing of distributed transactions in timestamping protocol in DDB.

The use of the mechanisms for transactions division and submission of certain values and their parameters makes the receiving of results from the execution of transactions in systems with distributed databases possible.

It is necessary to verify the developed algorithm modeling timestamp ordering in DDB and the comparison of the received results for throughput, response time and probability of service with the models of Two-version 2PL and 2PL with timestamps.

The created simulation model describes the real processes with a sufficient accuracy and allow receiving a reliable estimate for the changes of the throughput capability of the system in given parameters of the incoming transaction streams. The model is limited for the number of input flows and length of processed transactions because of the limitation of other similar models compared to them.

The conducted simulations and the results confirm the functionality of the modeling algorithm.

It is necessary to develop simulation models of concurrency control algorithms with much complexity by the meaning of the number of element replicas and transaction length in number of elements.

References

1. Connolly, T.M., Begg, C.E.: Database systems: Addison-Wesley (2002)
2. Date, C.J.: Introduction to Database Systems. 7th edn. Reading, MA: Addison-Wesley (2000)
3. Rozenkratz D., R. Stearns, P. Lewis II. System level concurrency control for distributed data base systems. ACM Trans. Database Systems, 3(2), 1978, p.178-198
4. TPC – C Benchmark. http://www.tpc.org/tpcc/spec/tpcc_current.pdf. (2007)
5. Vasileva, S.: Modeling of Primary Copy Two-Version Two Phase Locking. In: 4th International Conference on Information Systems and Grid Technologies, pp. 79—92, Sofia, May, 28-29. (2010)
6. Vasileva, S.: Algorithm for Primary Copy Locking with Timestamp Ordering. In: 5th International Conference on Information Systems and Grid Technologies, pp. 236—247, Sofia, May, 27-28. (2011)
7. Vasileva, S., Milev, A.: Simulation Models of Two-Phase Locking of Distributed transactions. In: International Conference on Computer Systems and Technologies, pp. V.12-1-V.12-6, ACM, New York (2008)
8. Таненбаум, Э., М. Стеен Расспределенные системы Принципы и парадигмы, Питер, Санкт Петербург, 2003.
9. Хасанова, Н., Разработка алгоритмов управления транзакциями, основанных на методе временных меток в распределенных базах данных, 2004, http://science.az/autoreferats/referat_hasanova_nazli.pdf.

* This paper supported by Project of Shumen University “Bishop Konstantin Preslavski” the topic of which is Information Models in Education and Business

Anycast DNS System in AS5421

Hristo Dragolov[1], Vesselin Kolev1[2], Stefan Dimitrov[3]

Associate at University Computer Centre, Sofia University, 5, J. Bourchier, Ave., 1164 Sofia, Bulgaria, hdragolov@ucc.uni-sofia.bg
[2]Researcher at TECHNION Israel Institute of Technology, Haifa, Israel, vlk@lcpe.uni-sofia.bg
[3]Assoc. Prof at Faculty of Mathematics and Informatics, Sofia University, 5, J. Bourchier, Ave., 1164 Sofia, Bulgaria, stefansd@fmi.uni-sofia.bg

Abstract. Anycast DNS system achieving three basic targets – reliability, security and manageability, is presented. The anycast system applied achieves high reliability by load balancing and redundancy. TSIG authentication and DNSSEC domain signing contribute to the high level of security. The implementation of a single stealth server from which all updates are downloaded and the application of DNSSEC-Tools set of software tools for DNSSEC domain signing facilitates management and maintenance of the system.

Keywords: DNS, anycast routing, BGP4/4+, TSIG, DNSSEC, stealth server, authoritative server, slave server, AS5421.

1 Introduction

The Anycast DNS system implemented is aimed at the creation of a robust and geographically dispersed platform that handles the inbound and outbound DNS traffic to a corporate network. The system should guarantee in case of AC power blackouts or breakout in communication line(s) to the network and services center the availability of cache (recursive) DNS service for the clients of the network, and the rest of the Internet users to be able to access the network resources (e.g. the web site www.abc.com), because the information for the domain abc.com and its subdomains remains accessible.

High level of security is achieved because a separate network segment and a VLAN is assigned to each DNS system. Additional level of security is introduced by the Transaction SIGnature (TSIG) authentication between the two types of servers in case of zone updates and transfers [1, 2], and by DNSSEC signing of domains [3, 4]. Failure of a given DNS server due to hacking will not disturb the normal operation of the rest of the DNS systems because all the recursive and authoritative server are mutually exchangeable.

The system described in the present paper is easily manageable. Changes in the domain are done on one and the same server (the stealth server). From there the other servers transfer updates for their data bases. DNSSEC signing and changes of keys is carried out easily with the help of the tools from the dnssec-tools package, that in practice automate the process [5].

2 DNS Anycast Overview

Anycast is based on the usage of routing and addressing policies to affect the most efficient path between a single source and several geographically dispersed targets that "listen" to a service within a receiver group. In Anycast, the same IP address space is used to address each of the listening targets (e.g. DNS servers). Layer 3 routing dynamically handles the calculation and transmission of packets from a given source (DNS Client) to its most appropriate target (DNS Server) [6, 7, 8].

In Figure 1 is shown one simple case of Anycast DNS. A single DNS client workstation, configured with the Anycast DNS IP address of 10.10.10.10, is shown performing DNS resolution against its "closest" of three DNS name servers deployed using the same Anycast IP address.

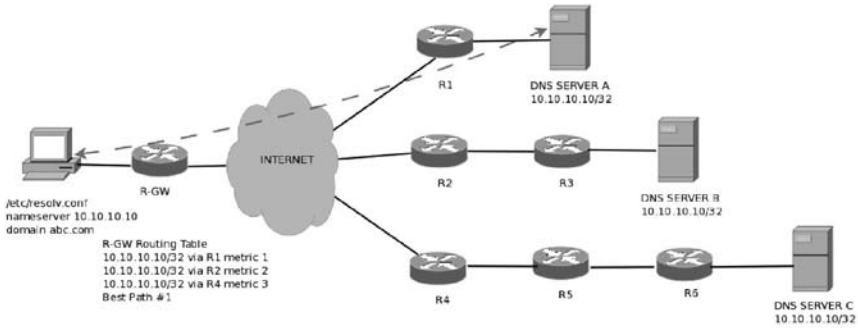


Fig. 1. A simple case of Anycast DNS.

The client's DNS resolver can resolve against any one of the three DNS servers. According to the network topology the client's gateway router (R-GW) would send the DNS client's trquest packets through router R1. Should router R1 or Server A fail, DNS client's packets would automatically be rerouted to the next nearest DNS server via routers R2 and R3, and so forth. Additionally, the route to server A, would be removed from the routing tables, thus preventing further use of that nameserver. Server A won't be used until it is restored and the IP Anycast address routes reinjected to the network.

3 The basic requirements for Anycast DNS

The basic set of requirements and recommendations for supporting Anycast DNS are as follows:

- ✓ Injection of Anycast IP address(es) into the routed network - This can be accomplished using either static routes or using dynamic routing protocols such as RIP, OSPF, or BGP [9].
- ✓ Host-based routing software that supports one of the major routing protocols

such as Quagga Routing Software [10].

- ✓ Clients should be configured to resolve DNS queries via the Anycast address(es).
- ✓ Nameservers should listen to DNS requests on the Anycast IP addresses.
- ✓ Nameservers should be configured with at least one Anycast IP address on a virtual (loopback or dummy) interface. Additionally, the server should be configured with a management IP which can be either a physical or an additional virtual interface.
- ✓ At least one physical IP (service IP) must be defined for the exchange of routing information, as well as, system access and maintenance in the absence of the routes to the Anycast IP address(es).
- ✓ Nameservers should be configured to use the physical or management IP addresses for zone-transfers, zone updates, and/or query-source because replies might go to a different server than intended.

4 Anycast Routing System Implementation

In the Anycast DNS system considered in the current paper BGP4/4+ protocol is used [11, 12]. The DNS system is thrice replicated, in each of the three basic campuses of the network (Figure 2).

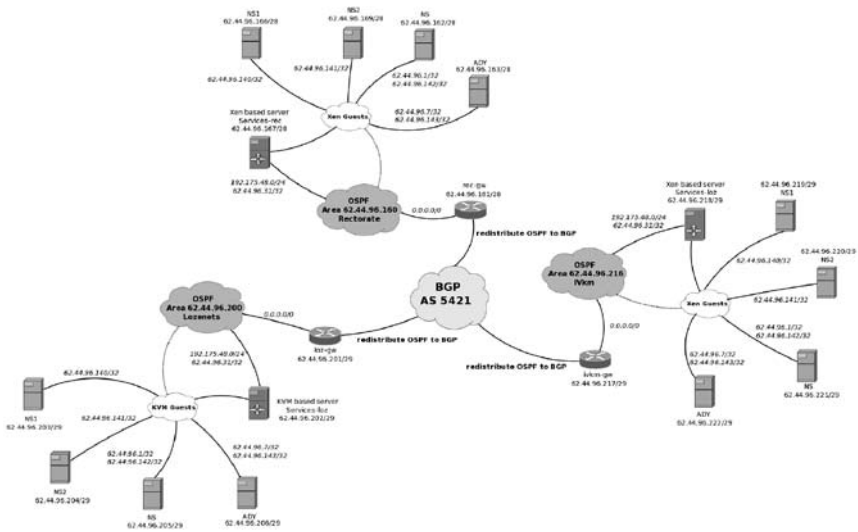


Fig. 2. The Anycast DNS system thrice replicated.

During normal operation (no connectivity failures are encountered), clients' requests within the the network are processed by the nearest node. For instance, a request originating from a computer in campus Rectorate are forwarded by

the main (gateway) router of the campus (REC-GW) to the anycast node in the same campus. That is, to the replica of the recursive server ns.uni-sofia.bg in this campus (62.44.96.142):

```
[root@ns ~]# traceroute ns.uni-sofia.bg
traceroute to ns.uni-sofia.bg (62.44.96.142), 30 hops max, 40 byte packets
 1  rec-gw.uni-sofia.bg (62.44.110.3)  0.100 ms  0.079 ms  0.072 ms
 2  ns.uni-sofia.bg (62.44.96.142)  0.648 ms  0.615 ms  0.613 ms
```

Requests for resource records from the zone of the domain uni-sofia.bg its subdomains and the respective in-addr.arpa and ip6.arpa zones originating from name servers outside the network, are processed by the main node of the network (in campus Lozenets). In case this node fails due to AC power blackout or break in connectivity, requests from external servers will be forwarded to the next in importance node (in campus Rectorate). If it fails the requests will be forwarded to the third node (in campus IV km). This hierarchy is achieved by manipulating the Local Preference attribute of the BGP protocol (Figure 3):

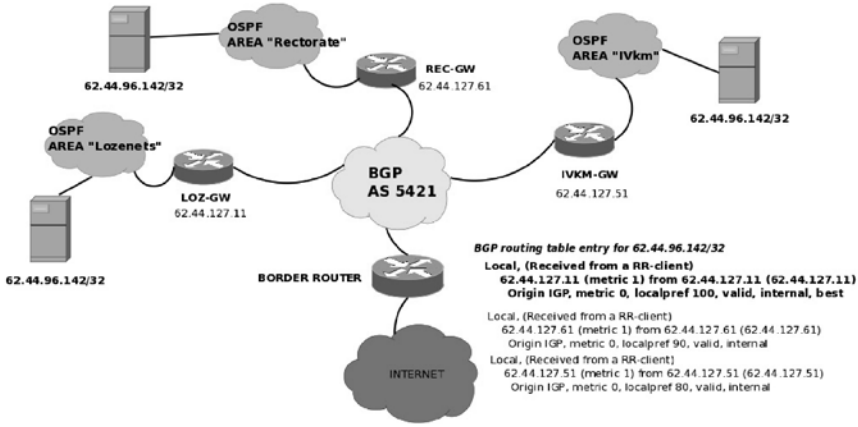


Fig. 3. Reachability of anycast cashing server from a border router.

Each of the three campuses has an independent connection to the Internet, providing for redundancy in processing of requests for resource records from the domain uni-sofia.bg, its subdomains and the respective in-addr.arpa and ip6.arpa domains.

5 Communication between Anycast Nodes within a Campus

The anycast DNS system in each campus is built on virtual technology - Xen or KVM [13, 14]. The kernel of each system is dom0, the machine that hosts the virtual “guests”. Each dom0 hosts four guests’ virtual machines (VMs): two of

them for the authoritive service, and the other two for the recursive (caching) service.

To the DNS segment a separate VLAN is assigned. The main (gateway) router for the respective campus (loz-gw, rec-gw and ivkm-gw) controls the routing process. The routing protocols OSPFv2 for IPv4 and OSPFv3 for IPv6 are used [15, 16]. Each of the latter routers serves as a default gateway sending a static route to all the participants in the respective OSPF area.

Each of the virual guests advertises in its OSPF area two network addresses: one service address and one anycast address. The latter address being the address of a dummy interface with /32 prefix. In its routing table each participant in the OSPF area holds paths to the service and anycast addresses of the rest of the participants as well as a default route, advertised by the gateway router. The full OSPF table is calculated at each of the participants, and at the main (gateway) router, as well. The latter redistributes the full OSPF table into the BGP protocol, too. In this way the border routers for the autonomous system (AS) are able to maintain redundancy on each of the DNS segments.

6 Profiling of the DNS Service

The DNS service is profiled [17, 18]. This means that the name servers that are authoritative for the domains' zones (i.e. the authoritative servers) are separated from those that process clients' requests to resolve names of resources on the Global network (i.e. the recursive/cache servers). Such a separation provides for the load balancing of the physical systems and the security of the DNS service.

Concerning load, the performance of the authoritative name servers should not depend on the load of the cache service. Cache servers very frequently become overloaded with requests (sometimes malicious) and their response time can be seriously downgraded. So they are resource (processor speed and memory capacity) greedy.

Security problems more strongly provide for the separation of authoritative service from the caching one. A major one being caching of malicious records by the recursive service which then can be returned as an answer to a request for a zone presumed as serviceable by the authoritative service. These are records that the server returns to clients in the so called "additional section" of its answer. Another problem is posed by abandoned and/or nonoperational zones when a server operates both as an authoritative and as a caching one. Clients requesting the caching service for records from a zone that is served by the authoritative service will receive answers directly from latter's zone file because its copy resides on its local disk. But, considering the case when the zone file is transferred from another (master) server and, for some reason, the transfer is broken for quite a long time, quite enough that the zone to be declared "expired". Error will be returned on client's request for a record from such a zone.

Similar problem will arise if the administrator of the zone decides not to use any more a given slave server and ceases the zone transfer to it. The administrator of the slave server that is a caching server too is not informed about the decision of the administrator of the master server. In this case the zone will enter in "expire" mode and clients' requests to it will not be answered. This means that clients will not see the domain defined in the zone file.

The above case is quite peculiar, so it should be illustrated with an example (Figure 4). A server (192.168.1.2) is authoritative for the domain example.com and as a "stealth" server for this zone it carries out regular transfers from its "master" server (192.168.1.1). In the same time the server 192.168.1.2 responds to clients' requests as a caching server, and answers to requests for www.example.com are taken directly from its zone file. In a certain moment the administrator of the zone of the domain example.com decides not to use server 192.168.1.2 as a "slave" any more, but misses to notify its administrator to cease servicing the zone of example.com. After the defined period of time with no transfer the name server 192.168.1.2 announces the zone "expired" and ceases to serve it. After that, if a client requests a record from that zone (e.g. www.example.com), it answers with error message. In the same time, clients using any other caching server but 192.168.1.2 have a normal access to all the records of example.com, because the server they request is singly and only a caching server.

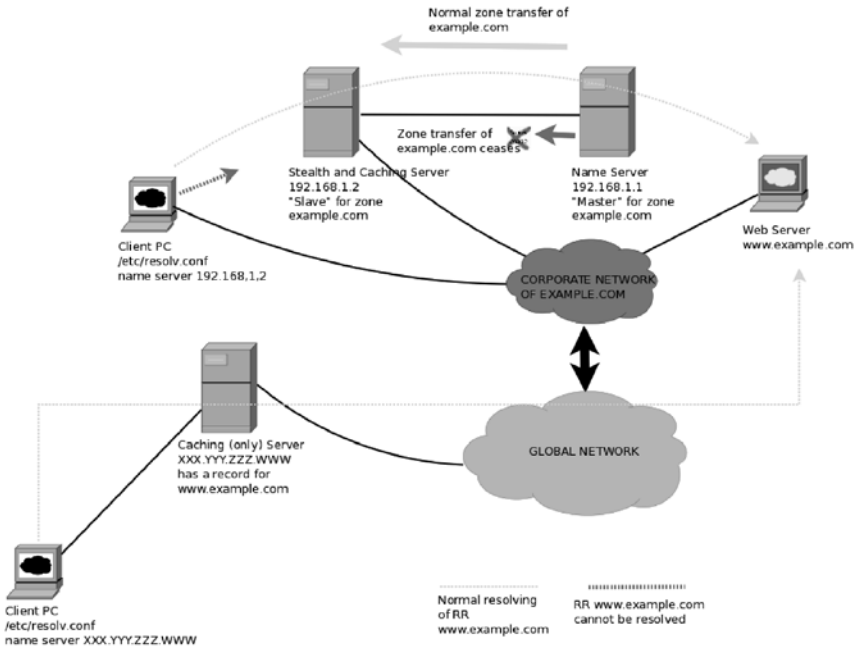


Fig. 4. The administrator of the zone decides not to use any more a slave server.

6.1 Authoritative Name Servers

The authoritative servers are responsible for the control of a given internet domain and delegate authoritative rights to other authoritative servers for the control of its subdomains. In our case the authoritative servers are responsible for the domain uni-sofia.bg, its subdomains and some of the respective reverse zones.

The authoritative servers are able to act as master and as slave servers providing redundancy and load balancing in this way. In the DNS infrastructure presented both authoritative servers act as slave servers. That means, the information about the domains (zones) they are authorised for, should be transferred from another server acting as master, i.e the stealth server situated in the main campus.

The authoritative servers within our Anycast DNS system are:

ns1.uni-sofia.bg (IPv4:62.44.96.140, IPv6:2001:67c:20d0:ff::140)
ns2.uni-sofia.bg (IPv4:62.44.96.141, IPv6:2001:67c:20d0:ff::141)

These name servers are included as NS resource records in the zone of the domain uni-sofia.bg and in the information concerning its delegation to the registrar of the ccTLD domain .bg - Register.BG [19].

6.2 Caching Name Servers

The caching name servers respond to clients' requests. Their IPv4 and/or IPv6 addresses are written into the file /etc/resolv.conf of UNIX/Linux system or into the DNS panel of an Windows station. The respective IPs are included in the host configuration delivered to clients by the Dynamic Host Configuration Protocol (DHCP) [20].

The caching name servers in the Anycast DNS system are:

ns.uni-sofia.bg (IPv4:62.44.96.142, IPv6:2001:67c:20d0:ff::142)
ady.uni-sofia.bg (IPv4:62.44.96.143, IPv6:2001:67c:20d0:ff::143)

These servers answer with higher priority requests from clients from the university network. But, in the name of the good will, are open to the public internet space.

6.3 The Stealth Server

As it was mentioned above, each slave authoritative server "takes" a particular zone of a domain from the basic master stealth server hosted in dom0 of the DNS segment in the main campus of the network. This server is of critical importance, so its availability is not known to anybody in global internet space.

All zones of domains and subdomains are stored on it. Any change in a domain is done in the domain's zone on that server. The latter advertises the change to all slave authoritative servers in all DNS segments by sending notify to them, authenticating the traffic by TSIG.

Name servers that are authoritative and master as well for certain zones and their respective reverse zones, i.e. they store the master copies of the zones on their local disks, are responsible to send notify to the stealth server (with the help of the option also-notify). The latter traffic is TSIG authenticated too. The stealth server on its behalf is responsible to send the changes to all authoritative slave servers (again with the help of the option also-notify), because they are set to receive updates from it.

On the stealth server is carried out DNSSEC signing and change of keys for the zone of the main domain and the respective reverse zone. Then notify is sent to all slave servers. Precise timing and synchronization for all servers is of great importance since the TSIG protocol uses timestamp on generating the hash to prevent replay of stored answers to requests. The Network Timing Protocol (NTP) is used for this purpose and is set on each DNS segment's dom0. While virtual guests use the mechanisms for clock setting provided by the virtual software [21].

7 Conclusion

The Anycast DNS system described is implemented in the network of the University of Sofia, Bulgaria. The network is well known to the Internet community as autonomous system (AS) 5421. The domain of the University of Sofia and the respective zone file is uni-sofia.bg and the reverse zone is 96.44.62.in-addr.arpa. The main campus of the network is Lozenets where the main control center is situated. The second in importance is Rectorate then comes IV km.

References

1. Secret Key Transaction Authentication for DNS (TSIG), http://www.ietf.org/rfc/rfc2845.txt
2. Използване на TSIG при комуникацията "клиент-сървър" в DNS, http://vesselin.org/papers/xhtml/bind9-tsig-roadwarriors.html (in Bulgarian)
3. DNSSEC: DNS Security Extensions, http://www.dnssec.net/
4. DNSSEC и BIND9 - кратко ръководство, http://vesselin.org/papers/xhtml/bind9-dnssec.html (in Bulgarian)
5. DNSSEC-Tools project, http://www.dnssec-tools.org/
6. Operation of Anycast Services, http://tools.ietf.org/pdf/rfc4786.pdf
7. Anycast DNS, http://www.netlinxinc.com/netlinx-blog/45/118.html
8. Reserved IPv6 Subnet Anycast Addresses, http://www.ietf.org/rfc/rfc2526.txt
9. Dynamic Routing, http://www.comptechdoc.org/independent/networking/guide/netdynamicroute.html
10. Quagga Routing Suite, http://www.nongnu.org/quagga/

11. Устройство на DNS anycast системата на СУ “Св. Климент Охридски”, https://mailbox.uni-sofia.bg/trac/wiki/DocumentationDNSAnycast (in Bulgarian)
12. A Border Gateway Protocol 4 (BGP-4), http://www.ietf.org/rfc/rfc1771.txt
13. http://www.xen.org/
14. Kernel Based Virtual Machine, http://www.linux-kvm.org/page/Main_Page
15. OSPF Version 2, http://www.ietf.org/rfc/rfc2328.txt
16. OSPF for IPv6, http://tools.ietf.org/html/rfc5340
17. DOMAIN NAMES - CONCEPTS AND FACILITIES, http://www.ietf.org/rfc/rfc1034.txt
18. Ron Aitchison, DNS and BIND 10, published by Apress
19. Domain Name System Structure and Delegation, http://tools.ietf.org/html/rfc1591#section-4
20. Dynamic Host Configuration Protocol, http://www.ietf.org/rfc/rfc2131.txt
21. Network Time Protocol Version 4: Protocol and Algorithms Specification, http://www.ietf.org/rfc/rfc5905.txt

Author Index

Airchinnigh, Mícheál Mac an 145, 163
Avdjieva, Irena 249
Blazeska-Tabakoska, Natasha 188, 206
Dimeski, Branko 44
Dimitrov, Stefan 365
Dimitrov, Vladimir 90, 139
Dragolov, Hristo 365
Georgieva-Trifonova, Tsvetanka 114
Goranov, Goran 38
Grigorova, Evgeniya 346
Hinova, Elena 38
Hristov, Hristo 98
Hristova, Radoslava 38, 294
Kaloyanova, Kalinka 55, 79, 280
Kjurchievska, Daniela 213
Kolev, Vesselin 365
Kovacheva, Zlatinka 11
Krachounov, Milko 173, 249, 302
Kulev, Ognyan 173, 249
Kyurkchiev, Hristo 79
Manev, Krassimir 326
Maneva, Neli 220
Manevska, Violeta 188, 206
Miceska, Elena 154
Mitreva, Emanuela 55
Mitrevski, Pece 240
Mufa, Vesna 188
Natchev, Nikolay 173
Nedelkoska, Jasmina 125
Nisheva-Pavlova, Maria 173, 231, 249
Pashov, Georgi 280
Pavlov, Pavel I. 28
Petrov, Peter 173, 249, 302
Peychev, Deyan 249
Popov, Ivan 302
Prisaganec, Milco 240
Rendevski, Nikola 240
Savoska, Snezana 44, 65
Semerdzhiev, Atanas 94, 134
Shiyachki, Dimitar 249
Shukerov, Dicho 196
Simeonova, Valeria 249
Stanev, Ivan 22
Tipografov, Milko 346
Todorova, Magdalina V. 312
Todorovska, Elena 249
Trifonov, Trifon 326
Vasileva, Svetlana 352
Vassilev, Dimitar 173, 249, 302
Zhelev, Radko 338