

Annals of Computer Science and Information Systems
Volume 17

Communication Papers of the 2018
Federated Conference on Computer Science
and Information Systems

September 9–12, 2018. Poznań, Poland



Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki (eds.)



Annals of Computer Science and Information Systems, Volume 17

Series editors:

Maria Ganzha (Editor-in-Chief),

Systems Research Institute Polish Academy of Sciences and Warsaw University of Technology, Poland

Leszek Maciaszek,

Wrocław University of Economy, Poland and Macquarie University, Australia

Marcin Paprzycki,

Systems Research Institute Polish Academy of Sciences and Management Academy, Poland

Senior Editorial Board:

Wil van der Aalst,

Department of Mathematics & Computer Science, Technische Universiteit Eindhoven (TU/e), Eindhoven, Netherlands

Marco Aiello,

Faculty of Mathematics and Natural Sciences, Distributed Systems, University of Groningen, Groningen, Netherlands

Mohammed Atiquzzaman,

School of Computer Science, University of Oklahoma, Norman, USA

Barrett Bryant,

Department of Computer Science and Engineering, University of North Texas, Denton, USA

Ana Fred,

Department of Electrical and Computer Engineering, Instituto Superior Técnico (IST—Technical University of Lisbon), Lisbon, Portugal

Janusz Górski,

Department of Software Engineering, Gdańsk University of Technology, Gdańsk, Poland

Giancarlo Guizzardi,

Free University of Bolzano-Bozen, Italy, Senior Member of the Ontology and Conceptual Modeling Research Group (NEMO), Brazil

Mike Hinchey,

Lero—the Irish Software Engineering Research Centre, University of Limerick, Ireland

Janusz Kacprzyk,

Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Irwin King,

The Chinese University of Hong Kong, Hong Kong

Juliusz L. Kulikowski,

Natęcz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland

Michael Luck,

Department of Informatics, King's College London, London, United Kingdom

Jan Madey,

Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw, Poland

Stan Matwin,

Dalhousie University, University of Ottawa, Canada and Institute of Computer Science, Polish Academy of Science, Poland

Marjan Mernik,

University of Maribor, Slovenia

Michael Segal,

Ben-Gurion University of the Negev, Israel

Andrzej Skowron,

Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw, Poland

John F. Sowa,

VivoMind Research, LLC, USA

Editorial Associates:

Katarzyna Wasielewska,

Systems Research Institute Polish Academy of Sciences, Poland

Paweł Sitek,

Kielce University of Technology, Kielce, Poland

T_EXnical editor: Aleksander Denisiuk,

University of Warmia and Mazury in Olsztyn, Poland

Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems

Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki
(eds.)

Annals of Computer Science and Information Systems, Volume 17
Communication Papers of the 2018 Federated Conference on Computer
Science and Information Systems

USB: ISBN 978-83-952357-1-9
WEB: ISBN 978-83-952357-0-2

ISSN 2300-5963
DOI 10.15439/978-83-952357-0-2

© 2018, Polskie Towarzystwo Informatyczne
Ul. Solec 38/103
00-394 Warsaw
Poland

Contact: secretariat@fedcsis.org
<http://annals-csis.org/>

Cover photo:
Aneta Tadra,
Elbląg, Poland

Also in this series:

Volume 16: Position Papers of the 2018 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-949419-8-7, ISBN USB: 978-83-949419-9-4**

Volume 15: Proceedings Papers of the 2018 Federated Conference on Computer Science and Information Systems, **ISBN Web 978-83-949419-5-6, ISBN USB 978-83-949419-6-3, ISBN ART 978-83-949419-7-0**

Volume 14: Proceedings of the First International Conference on Information Technology and Knowledge Management, **ISBN WEB: 978-83-949419-2-5, ISBN USB: 978-83-949419-1-8, ISBN ART: 978-83-949419-0-1**

Volume 13: Communication Papers of the 2017 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-922646-2-0, ISBN USB: 978-83-922646-3-7**

Volume 12: Position Papers of the 2017 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-922646-0-6, ISBN USB: 978-83-922646-1-3**

Volume 11: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-946253-7-5, ISBN USB: 978-83-946253-8-2, ISBN ART: 978-83-946253-9-9**

Volume 10: Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-65750-05-1, ISBN USB: 978-83-65750-06-8**

Volume 9: Position Papers of the 2016 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-93-4, ISBN USB: 978-83-60810-94-1**

Volume 8: Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-90-3, ISBN USB: 978-83-60810-91-0, ISBN ART: 978-83-60910-92-7**

Volume 7: Proceedings of the LQMR Workshop, **ISBN WEB: 978-83-60810-78-1, ISBN USB: 978-83-60810-79-8**

Volume 6: Position Papers of the 2015 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-76-7, ISBN USB: 978-83-60810-77-4**

DEAR Reader, it is our pleasure to present to you Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), which took place in Poznań, Poland, on September 9-12, 2018.

This year, communication papers were introduced for the second time, as a separate category of contributions. They report on research topics worthy of immediate communication. They may be used to mark a hot new research territory or to describe work in progress in order to quickly present it to scientific community. They may also contain additional information omitted from the earlier papers or may present software tools and products in a research state.

FedCSIS 2018 was Chaired by prof. Krzysztof Jassem, while dr. Paweł Skórzewski acted as the Chair of the Organizing Committee. This year, FedCSIS was organized by the Polish Information Processing Society (Mazovia Chapter), IEEE Poland Section Computer Society Chapter, Systems Research Institute Polish Academy of Sciences, Warsaw University of Technology, Wrocław University of Economics, and Adam Mickiewicz University.

FedCSIS 2018 was technically co-sponsored by: IEEE Region 8, IEEE Poland Section, IEEE Computer Society Technical Committee on Intelligent Informatics, IEEE Czechoslovakia Section Computer Society Chapter, IEEE Poland Section Gdańsk Computer Society Chapter, SMC Technical Committee on Computational Collective Intelligence, IEEE Poland Section Systems, Man, and Cybernetics Society Chapter, IEEE Poland Section Control System Society Chapter, IEEE Poland Section Computational Intelligence Society Chapter, ACM Special Interest Group on Applied Computing, International Federation for Information Processing, Committee of Computer Science of the Polish Academy of Sciences, Polish Operational and Systems Research Society, Mazovia Cluster ICT Poland and Eastern Cluster ICT Poland. FedCSIS 2018 was sponsored by Intel, Gambit, Samsung, Silver Bullet Labs, eSensei and Data Center PPNT.

During FedCSIS 2018, keynote lectures have been delivered by:

- Aksit, Mehmet, University of Twente, “*The Role of Computer Science and Software Technology in Organizing Universities for Industry 4.0 and Beyond*”
- Bosch, Jan, Chalmers University Technology, “*Towards a Digital Business Operating System*”
- Duch, Włodzisław, Nicolaus Copernicus University, “*Neurocognitive informatics for understanding brain functions*”
- O'Connor, Rory, V., Dublin City University, “*Demystifying the World of ICT Standardisation: An Insiders Viewpoint*”

FedCSIS 2018 consisted of the following events (conferences, symposia, workshops, special sessions). These events were grouped into FedCSIS conference areas, of various degree of integration. Specifically, those listed without indication of the year 2018 signify "abstract areas" with no direct paper submissions to them (but with submissions to their enclosed events).

- **AAIA'18 – 13th International Symposium Advances in Artificial Intelligence and Applications**
 - AIMaViG'18 – 3rd International Workshop on Artificial Intelligence in Machine Vision and Graphics
 - AIMA'18 – 8th International Workshop on Artificial Intelligence in Medical Applications
 - AIRIM'18 – 3rd International Workshop on AI aspects of Reasoning, Information, and Memory
 - ASIR'18 – 8th International Workshop on Advances in Semantic Information Retrieval
 - DMGATE'18 – 1st International Workshop on AI Methods in Data Mining Challenges
 - SEN-MAS'18 – 6th International Workshop on Smart Energy Networks & Multi-Agent Systems
 - WCO'18 – 11th International Workshop on Computational Optimization
- **CSS – Computer Science & Systems**
 - BEDA'18 – 1st International Workshop on Biomedical & Health Engineering and Data Analysis
 - CANA'18 – 11th Workshop on Computer Aspects of Numerical Algorithms
 - C&SS'18 – 5th International Conference on Cryptography and Security Systems
 - CPORA'18 – 3rd Workshop on Constraint Programming and Operation Research Applications
 - LTA'18 – 3rd International Workshop on Language Technologies and Applications
 - MMAP'18 – 11th International Symposium on Multimedia Applications and Processing
- **iNetSapp – International Conference on Innovative Network Systems and Applications**
 - INSERT'18 – 2nd International Conference on Security, Privacy, and Trust
 - IoT-ECAW'18 – 2nd Workshop on Internet of Things - Enablers, Challenges and Applications
- **IT4MBS – Information Technology for Management, Business & Society**
 - AITM'18 – 15th Conference on Advanced Information Technologies for Management
 - ISM'18 – 13th Conference on Information Systems Management
 - KAM'18 – 24th Conference on Knowledge Acquisition and Management
- **SSD&A – Software Systems Development & Applications**
 - MDASD'18 – 5th Workshop on Model Driven Approaches in System Development
 - MIDI'18 – 6th Conference on Multimedia, Interaction, Design and Innovation
 - LASD'18 – 2nd International Conference on Lean and Agile Software Development
 - SEW-38 & IWCP5-5 – Joint 38th IEEE Software Engineering Workshop (SEW-38) and 5th International Workshop on Cyber-Physical Systems (IWCP5-5)
- **DS-RAIT'18 – 5th Doctoral Symposium on Recent Advances in Information Technology**

Each paper, found in this volume, was refereed by at least two referees.

The program of FedCSIS required a dedicated effort of many people. Each event constituting FedCSIS had its own Organizing and Program Committee. We would like to express our warmest gratitude to all Committee members for

their hard work in attracting and later refereeing 394 regular submissions.

We thank the authors of papers for their great contribution to research and practice in computing and information systems. We thank the invited speakers for sharing their knowledge and wisdom with the participants. Finally, we thank all those responsible for staging the conference in Poznań. Organizing a conference of this scope and level could only be achieved by the collaborative effort of a highly capable team taking charge of such matters as conference registration system, finances, the venue, social events, catering, handling all sorts of individual requests from the authors, preparing the conference rooms, etc.

We hope you had an inspiring conference and an unforgettable stay in the beautiful city of Poznań. We also hope to meet you again for FedCSIS 2019 in Leipzig, Germany.

Co-Chairs of the FedCSIS Conference Series

Maria Ganzha, *Warsaw University of Technology, Poland and Systems Research Institute Polish Academy of Sciences, Warsaw, Poland*

Leszek Maciaszek, *Wrocław University of Economics, Wrocław, Poland and Macquarie University, Sydney, Australia*

Marcin Paprzycki, *Systems Research Institute Polish Academy of Sciences, Warsaw Poland and Management Academy, Warsaw, Poland*

Annals of Computer Science and Information Systems,
Volume 17

Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS)

September 9–12, 2018. Poznań, Poland

TABLE OF CONTENTS

13TH INTERNATIONAL SYMPOSIUM ADVANCES IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS

Call For Papers	1
Estimation of Intimacy Change in Team Using Vital Signs <i>Yuto Hattori, Tomoki Tanaka, Yusuke Kajiwara, Hiromitsu Shimakawa</i>	3

8TH INTERNATIONAL WORKSHOP ON ARTIFICIAL INTELLIGENCE IN MEDICAL APPLICATIONS

Call For Papers	11
Automated lung tumor detection and diagnosis in CT Scans using texture feature analysis and SVM <i>Tim Adams, Jens Dörpinghaus, Marc Jacobs, Volker Steinhage</i>	13

3RD INTERNATIONAL WORKSHOP ON AI ASPECTS OF REASONING, INFORMATION, AND MEMORY

Call For Papers	21
Mizar Set Comprehension in Isabelle Framework <i>Karol Pąk</i>	23
Inference rules for OWL-P in N3Logic <i>Dominik Tomaszuk</i>	27

8TH INTERNATIONAL WORKSHOP ON ADVANCES IN SEMANTIC INFORMATION RETRIEVAL

Call For Papers	35
A Chatbot Based On AIML Rules Extracted From Twitter Dialogues <i>Hiroshi Yamaguchi, Maxim Mozgovoy, Anna Danielewicz-Betz</i>	37

6TH INTERNATIONAL WORKSHOP ON SMART ENERGY NETWORKS & MULTI-AGENT SYSTEMS

Call For Papers	43
Design of models for the tokenization of electric power industry basing on the blockchain technology	45
<i>Artur Rot, Andrew Varnavskiy, Ulia Gruzina, Vladislav Trubnikov, Anastasia Buryakova, Ekaterina Sebechenko</i>	

11TH INTERNATIONAL WORKSHOP ON COMPUTATIONAL OPTIMIZATION

Call For Papers	51
”Passeport Vacances”: an assignment problem with cost balancing	53
<i>Corentin Beffa, Sacha Varone</i>	
Optimizing Maintenance in Project Management by Considering HSE and Resilience Engineering	61
<i>Behnam Einabadi, Pedram Memari, Seyed Farid Ghaderi</i>	

COMPUTER SCIENCE & SYSTEMS

Call For Papers	69
------------------------	-----------

1ST INTERNATIONAL WORKSHOP ON BIOMEDICAL & HEALTH ENGINEERING AND DATA ANALYSIS

Call For Papers	71
Assistive Smart, Structured 3D Environmental Information for the Visually Impaired and Blind: Leveraging the INSPEX Concept	73
<i>Suzanne Lesecq, Olivier Debicki, Laurent Ouvry, Christian Fabre, Nicolas Mareau, Julie Foucault, Francois Birot, Loic Sevrin, Steve Buckley, Carl Jackson, John Barrett, Alan McGibney, Susan Rea, David Rojas, Richard Banach, Joseph Razavi, Marc Correvoon, Gabriela Dudnik, Jean-Marc Van Gysegghem, Jean Herveg, Nathalie Grandjean, Florence Thiry, Cian O’Murchu, Alan Mathewson, Rosemary O’Keeffe, Andrea di Matteo, Vincenza Di Palma, Fabio Quaglia, Giuseppe Villa</i>	
Development of a mathematical model for electrode systems in rheophthalmography	83
<i>Petr Luzhnov, Anna Kiseleva, Dmitry Shamaev</i>	
Unintended effects of dependencies in source code on the flexibility of IT in organizations	87
<i>Deborah Tarenskeen, Rogier Van de Wetering, René Bakker</i>	
ECG signal coding methods in digital systems	95
<i>Tomasz Żentara, Krzysztof Murawski</i>	

11TH WORKSHOP ON COMPUTER ASPECTS OF NUMERICAL ALGORITHMS

Call For Papers	103
Benchmarking overlapping communication and computations with multiple streams for modern GPUs	105
<i>Paweł Czarnul</i>	
Accelerating Minimum Cost Polygon Triangulation Code with the TRACO Compiler	111
<i>Marek Pałkowski, Włodzimierz Bielecki</i>	
Parallelizing the code of the Fokker-Planck equation solution by stochastic approach in Julia programming language	115
<i>Anna Wawrzynczak</i>	

11TH INTERNATIONAL SYMPOSIUM ON MULTIMEDIA APPLICATIONS AND PROCESSING

Call For Papers	121
Developing keyword spotting method for the Polish language <i>Lukasz Laszko</i>	123
Soccer object motion recognition based on 3D convolutional neural networks <i>Jiwon Lee, Do-Won Nam, Wonyoung Yoo, Yoonhyung Kim, Minki Jeong, Changick Kim</i>	129
Analysis of inter-channel dependencies in audio lossless block coding <i>Cezary Wernik, Grzegorz Ulacha</i>	135

INTERNATIONAL CONFERENCE ON INNOVATIVE NETWORK SYSTEMS AND APPLICATIONS

Call For Papers	141
------------------------	------------

2ND WORKSHOP ON INTERNET OF THINGS - ENABLERS, CHALLENGES AND APPLICATIONS

Call For Papers	143
Adaptive Lighting System as a Smart Urban Object <i>Michael Aleithe, Philipp Skowron, Eric Schöne, Bogdan Franczyk</i>	145
Real Time Risk Monitoring in Fine-art with IoT Technology <i>Vincenza Carchiolo, Mark Phillip Loria, Marco Toja, Michele Malgeri</i>	151

INFORMATION TECHNOLOGY FOR MANAGEMENT, BUSINESS & SOCIETY

Call For Papers	159
------------------------	------------

16TH CONFERENCE ON ADVANCED INFORMATION TECHNOLOGIES FOR MANAGEMENT

Call For Papers	161
Business Process Management: Terms, Trends and Models <i>Renato Neder, Paulo Ramalho, Olivian Rabelo, Elisandra Zambra, Cristiano Maciel, Nathalia Benevides</i>	163
Development of crowd investing on the basis of ICO crypto assets using block-options for the supply of electric generation capacity <i>Artur Rot, Andrew Varnavskiy, Ulia Gruzina, Ekaterina Sebechenko, Vladislav Trubnikov, Anastasia Buryakova</i>	171
B2B Price Management using Price Waterfall Model and Business Intelligence solution <i>Krzysztof Senczyna, Radek Němec</i>	179

13TH CONFERENCE ON INFORMATION SYSTEMS MANAGEMENT

Call For Papers	187
A new task scheduling approach based on Spacing Multi-Objective Genetic algorithm in cloud <i>Kadda Beghdad Bey, Ali Belgacem, Hassina Nacer</i>	189
Data quality evaluation: a comparative analysis of company registers' open data in four European countries <i>Janis Bicevskis, Zane Bicevska, Anastasija Nikiforova, Ivo Oditis</i>	197
Evolution of the BPM Lifecycle <i>Marek Szelągowski</i>	205

<hr/>	
SOFTWARE SYSTEMS DEVELOPMENT & APPLICATIONS	
Call For Papers	213
<hr/>	
6TH CONFERENCE ON MULTIMEDIA, INTERACTION, DESIGN AND INNOVATION	
Call For Papers	215
Use of fuzzy cognitive maps for enhanced interaction with multiple mobile devices	217
<i>Przemysław Kucharski, Dawid Sielski, Tomasz Jaworski, Andrzej Romanowski, Jacek Kucharski</i>	
<hr/>	
JOINT 38TH IEEE SOFTWARE ENGINEERING WORKSHOP (SEW-38) AND 5TH INTERNATIONAL WORKSHOP ON CYBER-PHYSICAL SYSTEMS (IWCPS-5)	
Call For Papers	223
The Use of Gamification for Teaching Algorithms	225
<i>Luiz Ricardo Begosso, Luiz Carlos Begosso, Douglas Cunha, João Victor Pinto, Lucas Lemos, Michel Nunes</i>	
Automated generator for complex and realistic test data—a case study	233
<i>Richard Lipka, Tomas Potuzak</i>	
Assertional Reasoning for Concurrent and Communicating BPEL-like Programs	241
<i>Longfei Zhu, Qiwen Xu, Huibiao Zhu</i>	
Author Index	249

13th International Symposium Advances in Artificial Intelligence and Applications

A AIA'18 brings together scientists and practitioners to discuss their latest results and ideas in all areas of Artificial Intelligence. We hope that successful applications presented at AAIA'18 will be of interest to researchers who want to know about both theoretical advances and latest applied developments in AI.

TOPICS

Papers related to theories, methodologies, and applications in science and technology in the field of AI are especially solicited. Topics covering industrial applications and academic research are included, but not limited to:

- Decision Support
- Machine Learning
- Fuzzy Sets and Soft Computing
- Rough Sets and Approximate Reasoning
- Data Mining and Knowledge Discovery
- Data Modeling and Feature Engineering
- Data Integration and Information Fusion
- Hybrid and Hierarchical Intelligent Systems
- Neural Networks and Deep Learning
- Bayesian Networks and Bayesian Reasoning
- Case-based Reasoning and Similarity
- Web Mining and Social Networks
- Business Intelligence and Online Analytics
- Robotics and Cyber-Physical Systems
- AI-centered Systems and Large-Scale Applications

We also encourage researchers interested in the following topics to submit papers directly to the corresponding workshops, which are integral parts of AAIA'18:

- AI in Medical Applications (AIMA'18 workshop)
- AI in Machine Vision and Graphics (AIMaViG'18 workshop)
- AI in Reasoning Foundations (AIRIM'18 workshop)
- AI in Information Retrieval (ASIR'18 workshop)
- AI in Data Mining Challenges (DMGATE'18 workshop)
- AI in Smart Energy Networks (SEN-MAS'18 workshop)
- AI in Computational Optimization (WCO'18 workshop)

All submissions accepted to the main track of AAIA'18 and to the above workshops are treated equally in the conference programme and are equally considered for the paper awards.

PROFESSOR ZDZISŁAW PAWLAK BEST PAPER AWARDS

We are proud to continue the tradition started at the AAIA'06 and grant two "Professor Zdzisław Pawlak Best Paper Awards" for contributions which are outstanding in their scientific quality. The two award categories are:

- Best Student Paper. Papers qualifying for this award must be marked as "Student full paper" to be eligible.
- Best Paper Award.

Each award carries a prize of 300 EUR funded by the Mazowsze Chapter of the Polish Information Processing Society.

EVENT CHAIRS

- **Kwaśnicka, Halina**, Wrocław University of Science and Technology, Poland
- **Markowska-Kaczmar, Urszula**, Wrocław University of Science and Technology, Poland

ADVISORY BOARD

- **Kacprzyk, Janusz**, Polish Academy of Sciences, Poland
- **Marek, Victor**, University of Kentucky, United States
- **Matwin, Stan**, Dalhousie University, Canada
- **Michalewicz, Zbigniew**, University of Adelaide, Australia
- **Skowron, Andrzej**, University of Warsaw, Poland
- **Ślęzak, Dominik**, University of Warsaw, Poland

AREA SUPERVISORY COMMITTEE

- **Derksen, Christian**, SEN-MAS'18
- **Janusz, Andrzej**, DMGATE'18
- **Lasek, Piotr**, AIMA'18
- **Loukanova, Roussanka**, AIRIM'18
- **Markowska-Kaczmar, Urszula**, AAIA'18
- **Mozgovoy, Maxim**, ASIR'18
- **Śluzek, Andrzej**, AIMaViG'18
- **Zaharie, Daniela**, WCO'18

PROGRAM COMMITTEE

- **Baron, Grzegorz**
- **Bartkowiak, Anna**, Wrocław University, Poland
- **Bazan, Jan**, University of Rzeszów, Poland
- **Bembenik, Robert**
- **Betliński, Paweł**, Security On Demand, Poland
- **Błaszczyszki, Jerzy**, Poznań University of Technology, Poland
- **Chakraborty, Shampa**, Netaji Subhas Institute of Technology, India
- **do Carmo Nicoletti, Maria**, UFSCar & FACCAMP, Brazil
- **Franova, Marta**, CNRS, LRI & INRIA, France
- **Froelich, Wojciech**, University of Silesia, Poland
- **Gawrysiak, Piotr**
- **Girardi, Rosario**, UNIRIO, Brazil

- **Jaromczyk, Jerzy**, University of Kentucky, United States
- **Jatowt, Adam**, Kyoto University, Japan
- **Jin, Xiaolong**, Institute of Computing Technology, Chinese Academy of Sciences, China
- **Kasprzak, Włodzimierz**, Warsaw University of Technology, Poland
- **Korbicz, Józef**, University of Zielona Góra, Poland
- **Kryszkiewicz, Marzena**, Warsaw University of Technology, Poland
- **Kulikowski, Juliusz**, Institute of Biocybernetics and Biomedical Engineering, Poland
- **Lopes, Lucelene**, PUCRS, Brazil
- **Matson, Eric T.**, Purdue University, United States
- **Menasalvas, Ernestina**, Universidad Politécnica de Madrid, Spain
- **Miyamoto, Sadaaki**, University of Tsukuba, Japan
- **Moshkov, Mikhail**, King Abdullah University of Science and Technology, Saudi Arabia
- **Myszkowski, Paweł B.**, Wrocław University of Technology, Poland
- **Nowostawski, Mariusz**, Norwegian University of Technology and Science (NTNU), Norway
- **Ohsawa, Yukio**, University of Tokyo, Japan
- **Peters, Georg**, Munich University of Applied Sciences, Germany
- **Po, Laura**, Università di Modena e Reggio Emilia, Italy
- **Porta, Marco**, University of Pavia, Italy
- **Przybyła-Kasperek, Małgorzata**, University of Silesia, Poland
- **Raś, Zbigniew**, University of North Carolina at Charlotte, United States
- **Rauch, Jan**, University of Economics, Prague, Czech Republic
- **Reformat, Marek**, University of Alberta, Canada
- **Schaefer, Gerald**, Loughborough University, United Kingdom
- **Sikora, Marek**, Silesian University of Technology, Poland
- **Sikos, Leslie F.**, University of South Australia, Australia
- **Skonieczny, Lukasz**
- **Śluzek, Andrzej**, Khalifa University, United Arab Emirates
- **Stańczyk, Urszula**, Silesian University of Technology, Poland
- **Subbotin, Sergey**, Zaporizhzhya National Technical University, Ukraine
- **Sydow, Marcin**, Polish Academy of Sciences & Polish-Japanese Academy of Information Technology, Poland
- **Szczęch, Izabela**, Poznań University of Technology, Poland
- **Szczuka, Marcin**, University of Warsaw, Poland
- **Szpakowicz, Stan**, University of Ottawa, Canada
- **Szwed, Piotr**, AGH University of Science and Technology, Poland
- **Tomczyk, Arkadiusz**, Łódź University of Technology, Poland
- **Unland, Rainer**, Universität Duisburg-Essen, Germany
- **Unold, Olgierd**, Wrocław University of Technology, Poland
- **Zakrzewska, Danuta**, Łódź University of Technology, Poland
- **Zielosko, Beata**, University of Silesia, Poland
- **Ziółko, Bartosz**, AGH University of Science and Technology, Poland

Estimation of Intimacy Change in Team Using Vital Signs

Yuto Hattori, Tomoki Tanaka
Graduate School of Science and
Engineering Ritsumeikan University,
Shiga, Japan
Email:{yuto.hattori, tomoki}
@de.is.ritsumei.ac.jp

Yusuke Kajiwara
College of Production System Engineering
and Science Komatsu University,
Ishikawa, Japan
Email:yusuke.kajiwara@komatsu-u.ac.jp

Hiromitsu Shimakawa
College of Information Science and
Engineering Ritsumeikan University,
Shiga, Japan
Email:simakawa@cs.ritsumei.ac.jp

Abstract—In this research, we propose a method using vital signs, to estimate changes of the intimacy inside a team from interactions of team members in the same space. The method estimates both intimacy change between two members and that among whole members. The method facilitates team leaders to grasp the relationships and improve the team performance. Since various measurements representing features of the pulse wave are known to reflect personal emotion, we can expect to estimate the change of intimacy, providing the measurements with a machine learning algorithm. An experiment evaluating the proposed method showed high accuracy in the estimation among all members, but low accuracy in the estimation between two members. In both cases, the accuracy can be improved by choice of effective measurements. Through this experiment, we have found it is necessary to decide the effective measurements for each team to construct a model to estimate intimacy inside the team.

I. INTRODUCTION

TEAMS AIM to achieve common goals and objectives through information sharing and mutual support among interaction with other members. Team works can bring higher performance than personal efforts. However, there is a possibility to drop in performance of the team, because relationships between members get worse, or opinions are clashed because of differences in their thoughts and positions. To prevent and solve this problem, team leaders need to grasp relationships of team members at an early stage. If they succeed in it, they can improve the team performance [1].

However, teams consisting of many people form a number of relationships. The relationships change depending on the interaction between members. It is difficult for a team leader to grasp all relationships, which affects the team performance. Existing researches on human relationships analyze information exchanged among remote persons [2]–[6]. However, these tools may fail to reflect the actual interpersonal relationships. It is also suggested there may be privacy concerns in the analysis. In contrast, researches on human relationships in one space analyze conversation information transmitted to other people, but it is difficult to acquire the information in a real time manner. The record of the conversation might prevent team members from interacting as usual.

This paper proposes a method to estimate the change of the intimacy, which is an essential aspect of human relationships.

It examines two aspects: between two members and among whole members. For the estimation, we focus on an individual emotional state rather than information exchanged with others, because intimacy develops through arousing emotions. The pulse wave of each member is acquired by sensors in the course of team member interaction. Various feature quantities are calculated from the pulse wave indicating the emotional arousal. We estimate the change of the intimacy, providing the feature quantities with a machine learning algorithm. The estimated result makes it easy for team leaders to grasp relationships, which enable them to improve the team performance at an early stage.

II. ESTIMATION OF HUMAN RELATIONSHIPS WITH VITAL SIGNS

A. Estimation of human relationships

Existing research on estimation of human relationships have mainly used information in communication. Lin estimated relationships from the frequency of e-mails and mailing lists [2]. Garcia and Vigos et al. proposed a system which visualizes the number of exchanged messages and words that characterize one's correspondence with individuals, to estimate how the relationships have changed over time [3] [4]. Tago and He et al. used information in social media services to study of relationships [5] [6]. The method engages in estimating relationships in communication of people in different spaces. It is not available to estimate the relationships among people in one space such as a meeting.

Nishihara et al. examine the text of utterances between two persons [1]. They identify the role of the text from the combination of the particle and auxiliary verb in order to estimate the relationships. However, it costs high and consumes a lot of time to create the text of utterances that has been exchanged in a place such as a meeting. Furthermore, it is difficult to estimate the specific relationship when there are multiple relationships.

Many phonetic researches use the change of prosodic information such as pitch (fundamental frequency), speech rate and power [7]–[9]. However, it is difficult to acquire individual prosodic information in the place that many people get together.

B. The change of the personal emotion in interaction with many people

Though numerous works have studied detecting emotion in interaction with many people, there are few studies that have focused on human relationship. Hayamizu et al. acquired the facial expressions of group members with cameras, to estimate their emotions [9]. However, the devices which acquire feature information might prevent members from interacting as usual. Ohmoto et al. detected a social atmosphere of extrinsic involvement, enjoyment, or excitement from a body movement and vital signs with the aim to interact as usual [10]. However, it has been shown that expression of body movement varies with individuals in terms of degree of interaction, positions, and personality of individuals. It makes difficult to uniquely determine what body movement express. In addition, it has been shown that it is necessary to use vital signs, because expression of emotion is small when influenced by atmosphere.

C. Intimacy and emotion

A degree of intimacy affects teamwork such as communication and cooperativeness. Since individuals appropriately behave to maintain a stable relationship, the paper regards the intimacy as the quality of interactions between people [11]. Furthermore, the intimacy develops through the process where individuals disclose their personal information to others [12]. This personal information includes more thoughts or feelings rather than factual information. We focus on emotion in order to estimate the intimacy.

A degree of the intimacy increases, when emotions are transmitted through interaction. Besides, building of the intimacy relationship is promoted through arouses of emotions [13] [14]. Grasp of the change of emotional states would make it possible to estimate the change of the intimacy. Vital signs are often used to estimate the emotional state, because the change of emotional states would often appear on vital signs such as the pulse wave, the blood pressure, and the breathing pace.

D. The vital signs and emotion

In order to estimate the emotional state, HRV (Heart Rate Variability) is often used in the vital signs. Since HRV reflects variations in the balance of the sympathetic and the parasympathetic nerve that form autonomic nervous system, it is considered to reflect emotional state [15].

HRV is generally measured by an electrocardiogram. However, it has a disadvantage that its attachment restricts movement of a person. The existing researches suggested photoplethysmography as an alternative approach [16] [17]. It measures the fluctuation of the blood flow in the artery and the capillary by contraction of the heart. Eventually, it measures the pulse wave accompanying the heart rate. Peripheral sites densely packed with capillaries beneath the skin such as fingertips and earlobes are high in detection level. They are suitable to measure the pulse wave. The pulse wave sensor attached to the fingertip is detrimental to actions of a person wearing it, but earlobe causes no problem.

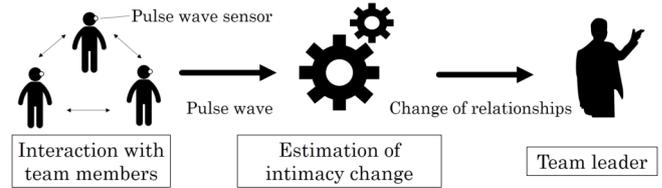


Fig. 1. Use case of method to estimate intimacy change

HR (Heart Rate) and RRI (R-R Interval), which is an index of HRV, are calculated from acquired pulse wave measurements. Activities of the sympathetic nerve brought by arousing emotions such as surprise and pleasure increase HR, while shorten RRI. In contrast, emotions such as relax make parasympathetic nerve active, which decreases HR and makes RRI longer. Furthermore, frequency analysis of RRI figures out the HF (High Frequency) part, and the LF (Low Frequency) part of it. The former ranges from 0.15 to 0.40 Hz, while the latter from 0.04 to 0.15 Hz. The HF part is used to quantify the parasympathetic nerve fluctuation, while the LF part indicates sympathetic and parasympathetic nerve fluctuation. Thus, the sympathetic nerve fluctuation is examined with LF/HF. HFnorm, which expresses the rate of HF components could be more powerful to reflect the sympathetic nerve fluctuation than HF [18]. These feature quantities are effective for estimating the change of the intimacy.

III. ESTIMATION OF INTIMACY CHANGE WITH VITAL SIGNS

A. Estimation method of the intimacy change

In this research, we propose the method using the pulse wave that is one of the vital signs, to estimate the change of intimacy inside a team. It is acquired with a pulse wave sensor attached on an earlobe of each team member. The method detects activities of the autonomic nerve. The more frequently the emotions are aroused, the larger the change in the activity of the sympathetic and the parasympathetic nerve configuring the autonomic nerve. When the degree of intimacy increases, the frequency of arousing emotions between team members is high, and change in activities of the autonomic nerves is large. On the contrary, when it decreases or does not changes, its activities hardly change. Therefore, the intimacy change of team members can be estimated from activities of the autonomic nerves.

Fig. 1. shows a use case of the method to estimate intimacy from the vital sign. Most of existing researches on human relationships focuses on those between the two people, but few researches focus on those among the whole team. In this research, we estimate not only the change of intimacy between two members but also that inside a team. The leader of team members can easily grasp the relationships among them from the initial stage after the team formation, using the estimation method. The method makes him improve the performance of the team.

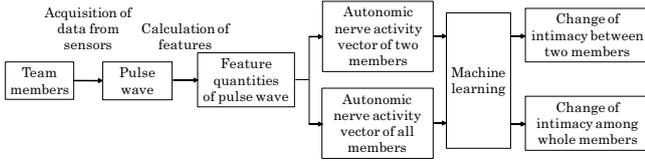


Fig. 2. The flow of the method to estimate change of intimacy

B. Construction of estimation model

Fig. 2. shows the flow of the method to estimate each change of intimacy. Pulse wave is measured using a sensor attached to each member during they join to an event. The method regards the event as a period of a team activity where they interact with each other. The pulse wave of each member is analyzed to figure out the feature quantities described in II.D.. Although a member influences other member, the autonomic nervous is not necessarily activated at the same time. We do not take changes in the course of the time into consideration. The activity vector of the autonomic nervous is composed of the mean and standard deviation of each feature quantities of a member i in the team is \mathbf{p}_i . A person who belongs to the team consisting of n members has relationships with $n - 1$ persons, where $n > 2$. It is not clear which member affects the change of the autonomic nervous of the person. Therefore, the intimacy change between specified two members (i, j) is estimated by providing $\mathbf{q}_{i,j}$, the activity vectors of autonomic nervous of the member i and j , determined from the following equation (1), to the machine learning algorithm.

$$\mathbf{q}_{i,j} = \mathbf{p}_i + \mathbf{p}_j \quad (1)$$

The intimacy change among whole members is estimated using these activity vectors of autonomic nervous \mathbf{r} , determined from the following equation (2) to the machine learning algorithm.

$$\mathbf{r} = \sum_{i=1}^n \mathbf{p}_i \quad (2)$$

We use a supervised learning. To constitute training data, the degree of intimacy evaluation of the other members is acquired two times by the questionnaire: before the event starts and after it ends. The difference these evaluations is regarded as the degree of intimacy change. The intimacy between two members and among whole members is classified based on the difference. This research defines 3 classes to estimate intimacy change: “Good Intimacy”, “Constant Intimacy” and “Bad Intimacy”, when the degree of the intimacy in questionnaire is increased, unchanged, and decreased, respectively. In the supervised machine learning, these three kinds of intimacy are adopted as objective variables, while the total activities of the autonomic nervous of target members are used as explanatory variables. The proposed method constructs an estimation model of intimacy change with training in the supervised learning.

IV. EXPERIMENT

A. Experiment overview

In this experiment, we acquired a pulse wave by the sensor when members are participating in the events. We evaluated possibility to estimate the intimacy change both of between two members and among whole members, using feature quantities acquired from the sensor.

Since the pulse wave differs depending on gender and age, subjects were selected 18 males who are undergraduate and graduate students. They were from 20 to 24 years old; whose average age is 21.4 years old. They are divided into 6 teams, each of which were constituted of 3 members. 3 teams consisted of strangers, while the others are acquaintances. We organized the stranger-teams, because they have no intimacy and tend to appear change of intimacy. In contrast, the acquaintance-teams already have established intimacy.

In each team, the change in the degree of the intimacy and the autonomic nervous is assumed to be different in early stage of human relationships. We explored the differences in each event.

Each subject wore a pulse wave sensor which is “Vital Meter” manufactured by TAOS on the earlobe to acquire the pulse wave.

As shown in Table I, 6 teams worked on preliminary events to acquire the initial intimacy between members. Its content is finding solutions to current affairs through discussion between members. In the preliminary events, we measured the initial degree of the intimacy within the team. Next, they worked on events to increase the degree of the intimacy. The second events are composed of 4 phases (ib1-ib4), which aim to increase the intimacy gradually. We created unique contents while referring to the existing methods [19]. Subjects answered to questionnaire (seven-point answer scale) about their degree of the intimacy with other members after each phase [20].

B. Method of analysis

First, we analyze the change of subjective evaluation acquired from the questionnaire and feature quantities calculated from the pulse wave in the events. We use each feature quantities standardized by the following equation (3), taking into account that there are differences in feature quantities of each subject.

$$L_i(t) = \frac{(g_i(t) - \bar{g}_i)}{\sigma_i} \quad (3)$$

$$\sigma_i = \sqrt{\frac{\sum_t (g_i(t) - \bar{g}_i)^2}{N - 1}} \quad (4)$$

where $g_i(t)$ is a feature quantity in an arbitrary time, and \bar{g}_i is the average of each feature quantity of ib1 to ib4.

Next, we evaluate the estimation accuracy of the intimacy, with regard to both of the intimacy change between two members and that among all members, using machine learning. In the case of two members, we compared total evaluation in all the teams with separate evaluation of the stranger-teams

TABLE I
THE EXPERIMENTAL PROCEDURE

Phase name (time/minutes)	Content of phase
work(10)	Finding solutions to presented assignment with team members
ib1 (6)	Self-disclosure to make other members have interest in
ib2 (6)	Self-disclosure including game element to break the tension
ib3 (6)	Cooperation work to deepen relationships
ib4 (6)	Cooperation thinking to deepen relationships

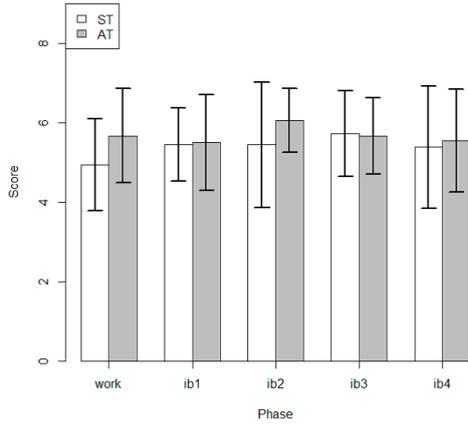


Fig. 3. The mean and standard deviation of questionnaire results

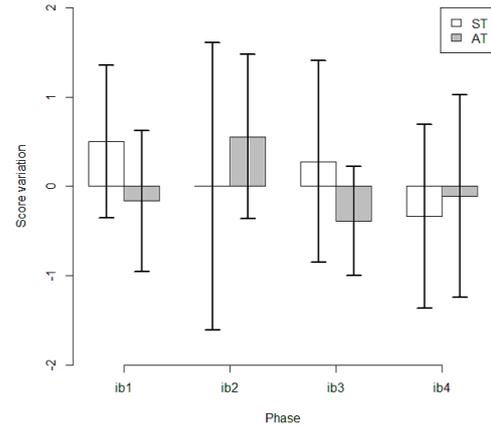


Fig. 4. The mean and standard deviation of the amount of change in the degree of intimacy with another member

and the acquaintance-teams in the experiment. In case of all three members, we evaluate only total evaluation in all the teams, because the sample size is small.

As a machine learning algorithm, we use the Random Forest, whose inputs are 12 variables composed of the activity vector of two members and all members. In addition, we also tried another estimation using only 3 important variables selected with the Random Forest importance measures. The one team is set test data and the others is training data. We evaluate the accuracy through 6 cross validation in estimation of all teams, while 3 cross validation in stranger-teams and acquaintance-teams. The evaluation index of the accuracy is F-measure determined from equation (5).

$$F - measure = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \quad (5)$$

C. Analysis of subjective evaluation

We analyze the subjective evaluation acquired from the questionnaire as the degree of the intimacy at event participation. Fig. 3. shows the mean and the standard deviation of each stranger-teams(ST) and acquaintance-teams(AT). No significant differences are observed.

Next, we focus on the degree of intimacy changes through each event. The amount of changes in the degree of intimacy with other members is determined from the following equation (6).

$$\Delta x_{ij,n} = x_{ij,n} - x_{ij,n-1} \quad (6)$$

where $x_{ij,n}$ represents the result of evaluation for the member (j) by the member (i) after ibn is finished ($1 \leq x_{i,j} \leq 7$). $\Delta x_{ij,n}$ is the amount of change in evaluation from the previous event, where $n = 0$ means the preliminary work. Fig. 4. shows the mean and the standard deviation of Δx in each team.

In stranger-teams, the degree of intimacy increased in $ib1$ and $ib3$, whereas decreased in $ib4$. In acquaintance-teams, the degree of intimacy decreased except in $ib2$. It represents that strangers have more feeling of closeness through the events. In addition, both of team members do not have a feeling of closeness in the phases but the difference in interaction with other members brings a difference in the degree of intimacy changes.

D. Analysis of feature quantities

We analyze the change of each feature quantity calculated from a pulse wave in each phase. The change in the mean of each feature quantity from $ib1$ to $ib4$ is calculated by the difference from the preliminary work. Fig. 5. and Fig. 6. show the changes in the mean of HR and RRI which reflect the autonomic nervous fluctuation in each phase of both stranger-teams and acquaintance-teams.

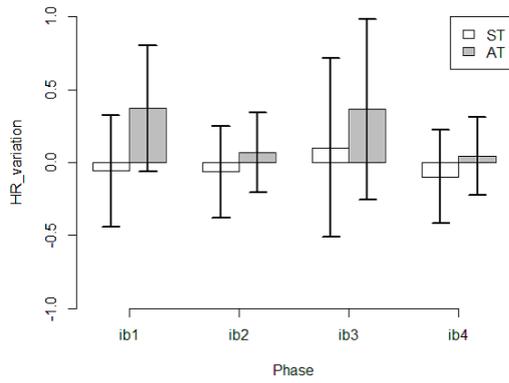


Fig. 5. The change in mean of HR

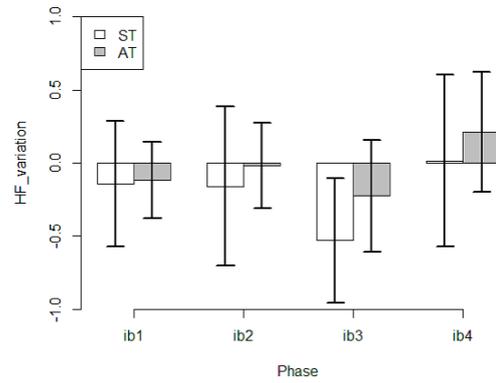


Fig. 7. The change in mean of HF

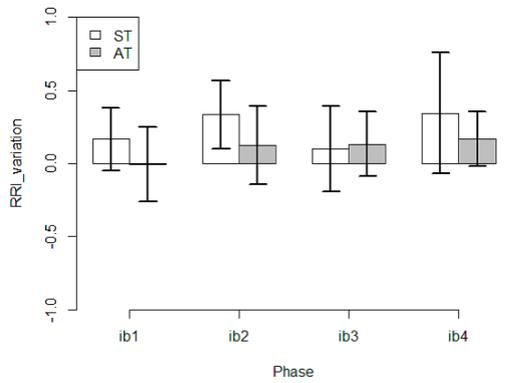


Fig. 6. The change in mean of RRI

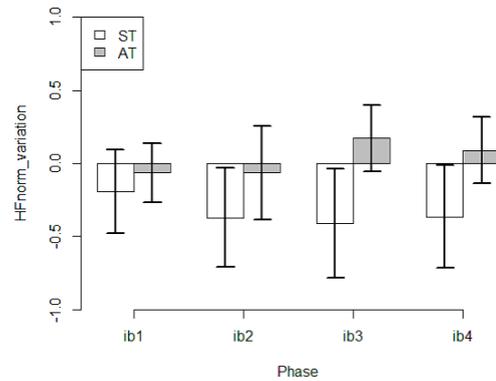


Fig. 8. The change in mean of HFnorm

In both stranger-teams and acquaintance-teams, the mean of HR was increased in ib3. In the stranger-teams, that in ib1, ib2 and ib4 was decreased. In contrast, in the acquaintance-terms, it was increased in all phases. The mean of RRI was increased except in ib1 of the acquaintance-teams.

Fig. 7. and Fig. 8. show the changes in the mean of HF and HFnorm which reflect the parasympathetic nerve fluctuation.

In both of the stranger-teams and the acquaintance-teams, the mean of HF was decreased from ib1 to ib3. In addition, the mean of HFnorm was decreased in all phases.

Finally, Fig. 9. and Fig. 10. show the changes in the mean of LF/HF which reflects sympathetic nerve fluctuation and that of LF which reflects sympathetic nerve and parasympathetic nerve fluctuation, respectively.

In both of the stranger-teams and the acquaintance-teams, the mean of LF/HF was increased from ib1 to ib3. In addition, the mean of LF was increased in all phases.

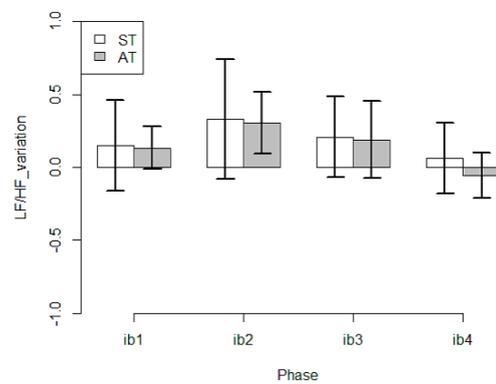


Fig. 9. The change in mean of LF/HF

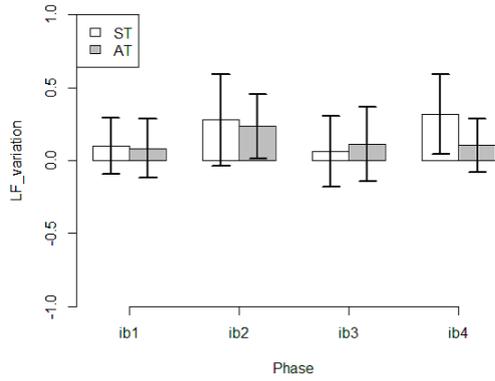


Fig. 10. The change in mean of LF

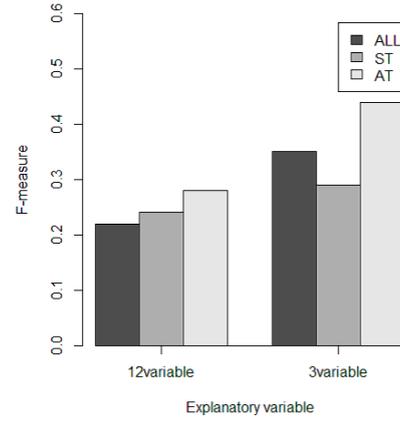


Fig. 11. Each estimation accuracy of intimacy between two members

TABLE II
THE SAMPLE SIZE OF 3 CLASSES: “GOOD INTIMACY”(GI), “CONSTANT INTIMACY”(CI) AND “BAD INTIMACY”(BI)

	GI	CI	BI
ALL	24	26	22
ST	14	13	9
AT	10	13	13

TABLE III
THE CROSS-TABULATION TABLE IN ESTIMATION OF ALL TEAMS

	BI	CI	GI
BI	278	870	691
CI	1186	888	1208
GI	736	842	501
Recall	0.12	0.34	0.21
Precision	0.15	0.3	0.22
F-measure	0.13	0.3	0.22

E. Estimation accuracy of intimacy between two members

We evaluate the estimation accuracy on three classes of “Good intimacy”, “Constant intimacy” and “Bad intimacy” from the changes of the intimacy between two members. The amount of change in the degree of intimacy between two members is determined from the following equation (7).

$$\Delta y_{ij,n} = x_{ij,n} + x_{ji,n} \quad (7)$$

where $\Delta y_{ij,n}$ is summing up amount of change from the previous phase in evaluation of members (i and j) after ibn is finished. $\Delta y_{ij,n}$ is classified into 3 classes in order to be adapted objective variable: $\Delta y_{ij,n} > 0$ (“Good Intimacy”), $\Delta y_{ij,n} = 0$ (“Constant Intimacy”) and $\Delta y_{ij,n} < 0$ (“Bad Intimacy”). Table II shows each sample size.

We estimated using not only both 12 variables, but also 3 variables which is the mean of HR, the mean of RRI, and the standard deviation of RRI as explanatory variable. Fig. 11. shows the estimation accuracy.

The estimation accuracy in case of using 12 variables was low, namely, it is lower than the accuracy of 1/3 which is random guess. In contrast, the estimation accuracy of all teams, stranger-teams and acquaintance-teams in case of using 3 variables was higher than that of 12 variables. The estimation accuracy of the acquaintance-teams was especially high, though that of the stranger-teams was lower than all teams. Table III shows cross-tabulation table in the estimation of all teams. The identification accuracy of “Good Intimacy”

and “Bad Intimacy” was low. It is misidentified as “Constant Intimacy”.

F. Estimation accuracy of intimacy among all members

We evaluate the estimation accuracy from 3 changes of the intimacy in a team. However, in this case, the sample size was too small to discriminate either of 3 classes or each of stranger-teams and acquaintance-teams. Because of this, we evaluate the estimation accuracy of 2 classes, one of which is “Good Intimacy” and the other is either of “Constant Intimacy” or “Bad Intimacy” in all teams. The amount of changes in the degree of intimacy among all members is determined from the following equation (8).

$$\Delta z_{ijk,n} = \Delta y_{ij,n} + \Delta y_{jk,n} + \Delta y_{ik,n} \quad (8)$$

where $\Delta z_{ijk,n}$ is the sum of the amount of 3 changes from the previous phase in evaluation of members, assuming k is same team members with i and j , after ibn is finished. $\Delta z_{ijk,n}$ is classified into 2 classes in order to be adapted objective variable: $\Delta z_{ijk,n} > 0$ (“Good Intimacy”) and $\Delta z_{ijk,n} \leq 0$ (“Constant Intimacy” and “Bad Intimacy”). The sample size of the two cases is 11 and 13, respectively. We estimated using not only while 12 variables, but also 3 variables which are the mean of HF, the mean of HFnorm and the standard deviation of HFnorm as explanatory variables. The estimation accuracy was 0.69 in case of 12 variable, and 0.76 in 3 variable. In

estimation among all members, both of the accuracy of 12 variables and 3 variables were high.

V. DISCUSSION

A. The change of intimacy and feature quantities in phase

The stranger-teams were expected that the degree of the intimacy increased in phases through events, while no change was expected in acquaintance-teams. However, the result was quite different; the degree of intimacy changes varies with events. We discuss reasons for this result below.

In ib1, it is assumed that strangers had feeling of closeness by learning about others. In contrast, acquaintances who could obtain only known information from others did not have more than original feeling of closeness. In ib2, acquaintances had feeling of closeness through sharing information while enjoying with others by games. On the other hand, strangers were unaffected with games. Their feeling was unchanged because it was similar to ib1, regarding self-disclosure. In ib3, strangers had feeling of closeness because of close relationships to others through cooperative works. In contrast, acquaintances worked with certain relationships, which made the degree of intimacy decreased from ib2. In ib4, both strangers and acquaintances did not have feeling of closeness because almost everyone concentrated on thinking than discussing among the members.

From HR of the feature quantities calculated from the pulse wave increased in ib3, it is assumed that sympathetic nerve got active, because of aroused emotions such as joy and pleasure coming from increasing opportunities to interact with other members. HR was decreased in ib1, ib2 and ib4 for the stranger-teams. However, HF was decreased in ib1 and ib2. It implies that it might be affected by either tension in works, or temporal change in breathing such as laughters in conversations.

It can be considered that there is a difference in autonomic nervous changes depending on the contents of the interaction. However, the change might vary with person to person. It is necessary to consider that the influence of the sympathetic and parasympathetic nerves on each person.

B. Estimation accuracy of intimacy change

As a result of estimation based on the features acquired from the pulse wave, it is possible to estimate the intimacy in team. However, estimation between two members is difficult. As the possible reason for this, the interaction of whole team seemed to be more frequent than that between two members. The interaction of whole team affected the emotional arousal of all members. Table IV shows the mean and standard deviation of 3 important variables in each class for estimation of whole team. Only the standard deviation of HFnorm was recognized significant difference in the results of U test whose significance level is 5%. It shows the variance of parasympathetic nerve activity of members who feel close each other is larger than other relationships. In the intimacy between two members, it is assumed that estimation accuracy is low because of many misidentified. It might come from the interaction between two

TABLE IV
THE MEAN AND STANDARD DEVIATION OF 3 VARIABLES IN 2 CLASSES

	HF_Mean		HFnorm_Mean		HFnorm_SD	
	Mean	SD	Mean	SD	Mean	SD
CI,BI	-0.07	1.59	-0.31	0.84	5.48	0.46
GI	0.09	1.03	0.36	0.9	6.17	0.44

members. Since it was little, we could not find difference in the activity vectors of autonomic nervous of two members.

In this experiment, since the degree of intimacy was acquired by subjective evaluation, the subjects may evaluate their intimacy increased temporary because of arousing emotions according to the peak-end rule [21]. However, activity vectors of autonomic nervous smooth temporary arousing emotions, which make the estimation accuracy low. The estimation accuracy might improve, if we consider the peak value of each feature quantity and the ratio of the peak over a threshold value in the whole using the Peak-Valley method. The estimation accuracy of both teams was improved by using only 3 important variables. However, the important variables were different in each team. In addition, since a sufficient sample size was not obtained, we cannot decide important feature quantities uniquely, to estimate the intimacy.

VI. CONCLUSION

In this paper, we proposed a method to estimate intimacy changes not only between two members but also among whole members using feature quantities calculated from the pulse wave, in order to grasp relationships of team members. From the results of our experiment, the change of intimacy degree was different between strangers and acquaintances, depending on the way of interaction. Individual differences in the effects of autonomic nervous were shown from changes of both parasympathetic and sympathetic. Estimation accuracy of changes in the intimacy among whole members was high enough to estimate relationships. However, the estimation accuracy between two members was low. The selection of only effective feature quantities improved the estimation accuracy in the both cases, but we cannot identify the feature quantities uniquely. It is necessary to select feature quantities for each team. Since the number of samples in this experiment is small, it is necessary to obtain accurate results by conducting an additional experiment and increasing the number of samples in the future.

Since the method examines the pulse wave, it can be applied only in situation where all members stay sitting, such as discussion. In this experiment, we assumed 2 kinds of teams: teams of strangers and teams of acquaintances. However, in actual organizations and companies, there can be teams in which both of them are mixed. It is necessary to examine whether the proposed method works well for such teams in future.

REFERENCES

- [1] Nishihara, Y., Sunayama, W., & Yachida, M. (2008). Human Friendship and Hierarchical Relationship Estimation from Utterance Texts, *IEICE transactions on information and systems*, 91.1: 78-88.
- [2] Lin, H. (2010, December). Predicting sensitive relationships from email corpus. In *Genetic and Evolutionary Computing (ICGEC), 2010 Fourth International Conference on* (pp. 264-267). IEEE.
- [3] Garcia, T., Aires, J., & Goncalves, D. (2012, May). Who have I been talking to?. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 481-484). ACM.
- [4] Vidas, F. B., Golder, S., & Donath, J. (2006, April). Visualizing email content: portraying relationships from conversational histories. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 979-988). ACM.
- [5] Tago, K., & Jin, Q. (2017, November). Analyzing influence of emotional tweets on user relationships by naive bayes classification and statistical tests. In *Service-Oriented Computing and Applications (SOCA), 2017 IEEE 10th International Conference on* (pp. 217-222). IEEE.
- [6] He, X., Wang, Y., Li, Y., & Jiang, Y. (2018, January). Investigating Relationships in Online Communities: A Social Network Analysis. In *Proceedings of the 2018 2nd International Conference on Management Engineering, Software Engineering and Service Sciences* (pp. 229-233). ACM.
- [7] Nishimura, R., Kitaoka, N., & Nakagawa, S. (2008). Analysis of relationship between impression of human-to-human conversations and prosodic change and its modeling. In *Ninth Annual Conference of the International Speech Communication Association*.
- [8] Inaguma, H., Inoue, K., Nakamura, S., Takahashi, K., & Kawahara, T. (2016, November). Prediction of ice-breaking between participants using prosodic features in the first meeting dialogue. In *Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction* (pp. 11-15). ACM.
- [9] Hayamizu, T., Mutsuo, S., Miyawaki, K., Mori, H., Nishiguchi, S., & Yamashita, N. (2012, November). Group emotion estimation using Bayesian network based on facial expression and prosodic information. In *Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on* (pp. 177-182). IEEE.
- [10] Ohmoto, Y., Miyake, T., & Nishida, T. (2010). Study on Detecting a Social Atmosphere of Extrinsic "Involvement, Enjoyment, or Excitement" in Multi-User Interaction, *IEICE transactions on information and systems*, 93.6: 870-878
- [11] Laurenceau, J. P., Barrett, L. F., & Pietromonaco, P. R. (1998). Intimacy as an interpersonal process: The importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges. *Journal of personality and social psychology*, 74(5), 1238.
- [12] Reis, H. T., & Shaver, P. (1988). Intimacy as an interpersonal process. *Handbook of personal relationships*, 24(3), 367-389.
- [13] Howell, A., & Conway, M. (1990). Perceived intimacy of expressed emotion. *The Journal of social psychology*, 130(4), 467-476.
- [14] Graham, S. M., Huang, J. Y., Clark, M. S., & Helgeson, V. S. (2008). The positives of negative emotions: Willingness to express negative emotions promotes relationships. *Personality and Social Psychology Bulletin*, 34(3), 394-406.
- [15] Task Force of the European Society of Cardiology. (1996). Heart rate variability, standards of measurement, physiological interpretation, and clinical use. *circulation*, 93, 1043-1065.
- [16] Senghipphany, T., Tretriluxana, S., & Chitsakul, K. (2015, June). Comparison of Heart Rate statistical parameters from Photoplethysmographic signal in resting and exercise conditions. In *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2015 12th International Conference on* (pp. 1-5). IEEE.
- [17] Bolanos, M., Nazeran, H., & Haltiwanger, E. (2006, August). Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE* (pp. 4289-4294). IEEE.
- [18] Yin, L., Zhang, W., & Xia, L. (2010, October). Effect of skin pressure caused by the cuff on female heart rate variability. In *Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference on* (Vol. 2, pp. 883-886). IEEE.
- [19] Hori, K., Kato, A., & Karube, T. (2007). Team Building: hito to hito wo "tunagu" gihou [Team Building: Technique to connect people]. Tokyo: Nikkei Publishing Inc.
- [20] Fujimori, T. (1980). EFFECTS OF ATTITUDE SIMILARITY AND TOPIC IMPORTANCE ON INTERPERSONAL ATTRACTION. *The Japanese Journal of Experimental Social Psychology*, 20(1), 35-43.
- [21] Miron-Shatz, T. (2009). Evaluating multi-episode events: Boundary conditions for the peak-end rule. *Emotion*, 9(2), 206.

8th International Workshop on Artificial Intelligence in Medical Applications

THE workshop on Artificial Intelligence in Medical Applications – AIMA'2018—provides an interdisciplinary forum for researchers and developers to present and discuss latest advances in research work as well as prototyped or fielded systems of applications of Artificial Intelligence in the wide and heterogeneous field of medicine, health care and surgery. The workshop covers the whole range of theoretical and practical aspects, technologies and systems based on Artificial Intelligence in the medical domain and aims to bring together specialists for exchanging ideas and promote fruitful discussions.

TOPICS

- Artificial Intelligence Techniques in Health Sciences
- Knowledge Management of Medical Data
- Data Mining and Knowledge Discovery in Medicine
- Health Care Information Systems
- Clinical Information Systems
- Agent Oriented Techniques in Medicine
- Medical Image Processing and Techniques
- Medical Expert Systems
- Diagnoses and Therapy Support Systems
- Biomedical Applications
- Applications of AI in Health Care and Surgery Systems
- Machine Learning-based Medical Systems
- Medical Data- and Knowledge Bases
- Neural Networks in Medicine
- Ontology and Medical Information

- Social Aspects of AI in Medicine
- Medical Signal and Image Processing and Techniques
- Ambient Intelligence and Pervasive Computing in Medicine and Health Care

EVENT CHAIRS

- **Paja, Wiesław**, University of Rzeszów, Poland
- **Pancerz, Krzysztof**, University of Rzeszów, Poland
- **Stocean, Catalin**, University of Craiova, Romania

PROGRAM COMMITTEE

- **Belciug, Smaranda**, University of Craiova, Romania
- **Iantovics, Barna**, Petru Maior University, Romania
- **Lasek, Piotr**, University of Rzeszow, Poland
- **Leniowska, Lucyna**, University of Rzeszow, Poland
- **Lichtblau, Daniel**, Wolfram Research, United States
- **Majernik, Jaroslav**, Pavol Jozef Safarik University in Kosice, Slovakia
- **Mapayi, Temitope**, University of KwaZulu-Natal, Durban, South Africa, South Africa
- **Olszewska, Joanna Isabelle**, University of Gloucestershire, United Kingdom
- **Perner, Petra**, IBAI Leipzig, Germany
- **Salem, Abdel-Badeeh M.**, Ain Shams University, Egypt
- **Stańczyk, Urszula**, Silesian University of Technology, Poland
- **Stocean, Ruxandra**, University of Craiova, Romania
- **Zaitseva, Elena**, University of Zilina, Slovakia

Automated lung tumor detection and diagnosis in CT Scans using texture feature analysis and SVM

Tim Adams^{*†}, Jens Dörpinghaus^{*}, Marc Jacobs^{*} and Volker Steinhage[†]

^{*}Fraunhofer Institute for Algorithms and Scientific Computing,
 Schloss Birlinghoven, Sankt Augustin, Germany

[†] Department of Computer Science, University of Bonn

[‡] Tim Adams has been with the Dept. of Computer Science, University of Bonn at the time of this work
 Email: tim.adams@scai.fraunhofer.de, jens.doerpinghaus@scai.fraunhofer.de, marc.jacobs@scai.fraunhofer.de,
 steinhage@cs.uni-bonn.de

Abstract—CT scans are an important tool in the diagnosis of lung tumors in medicine. This work presents an automated system for lung tumor diagnosis on CT scans. Scans are automatically segmented using marker-based watershed transformation, which successfully segments hardly separable, lung wall adjunct tumors. The scans are further analyzed in a sliding window approach using Haralick features and a Support Vector Machine classifier to detect and classify benign and malignant tumors. This novel approach for classification was tested using the LUNGx Challenge dataset [1] and achieved exceptional results while utilizing a minimal training set.

I. INTRODUCTION

CANCER is still one of the most frequent causes of death worldwide [2]. Lung cancer is in the course of this the leading cause of cancer deaths for men as well as one of the most common cancers diagnosed in woman [3]. An early diagnosis is important, as it can influence the choice of treatment and thus prolong the patient's life. A widely used diagnostic method is the analysis of computed tomography (CT) scans of the lung. Recent work has shown that selected texture features can be used on CT scans to distinguish between benign and malignant pulmonary nodules [4], [5], [6]. The presented approach introduces a workflow that automatically identifies and classifies tumor tissue in lung CT scans by extracting *Haralick texture features* [7] and classifying image regions in a sliding window approach using a *Support Vector Machine* (SVM) classifier.

II. BACKGROUND

For a systematic detection of tumors in lung CT scans several sub-problems have to be considered. We will discuss them together with their state of the art solutions. The general procedure and methodology of a tumor diagnosis system can be subdivided into the following four sub-problems [8]:

- 1) **Preprocessing:** The goal of preprocessing is the reduction of unwanted artifacts and noise often occurring in CT scans. The preprocessing step facilitates the further processing of the image and may also be used to enhance certain image features for later processing.
- 2) **Segmentation:** Segmentation is used to separate semantically coherent image areas. It is a crucial step in order to achieve a successful classification, because the

segmentation result significantly influences the results of the following processing steps.

- 3) **Feature Extraction:** This step uses algorithms to extract selected features from the image. Lung tumors often differ in size, texture or contour.
- 4) **Classification:** After the feature extraction, each identified region is evaluated based on its characteristics. Based on the rating of the chosen classifier, images or image areas may be assigned to a positive or negative class.

Based on these sub-steps a system for tumor recognition and classification can be created from a combination of different approaches that are capable of solving one or more of these subproblems. For segmentation, feature extraction and classification numerous different methods can be utilized. Recent works concentrating on texture features for cancer analysis archive promising classification results using an SVM classifier for tumor detection and evaluation. A recent work by Nilesh Bhaskarrao Bahadure et al. [9] shows the impact of texture features in combination with an SVM classifier for tumor detection in brain MRI scans. They achieve an accuracy value of 96.51%. The use of texture features for tumor detection has also been studied in the area of lung CT scans by several research groups:

- Zayed and Elnemr [4] study the effectiveness of *Haralick texture features* on the identification of lungs with malign pulmonary nodules. For segmentation, the lung with the largest volume is mirrored and used as a mask for the second lung to separate tumors inter-grown with the lung wall. They conclude that selected texture features could be useful for the detection of abnormalities in CT lung scans. Although this approach may detect abnormal lungs, the position of the abnormal tissue within the diseased lung cannot be determined.
- Han et al. [5] compare the performance of different descriptors (*Haralick 2D* and *3D*, *Gabor*, *Local Binary Patterns*) for classification of benign and malignant lung tumors. For training, they use CT scans with outer tumor borders annotated by different radiologists. Han et al. discuss that *Haralick features* yield the best classification

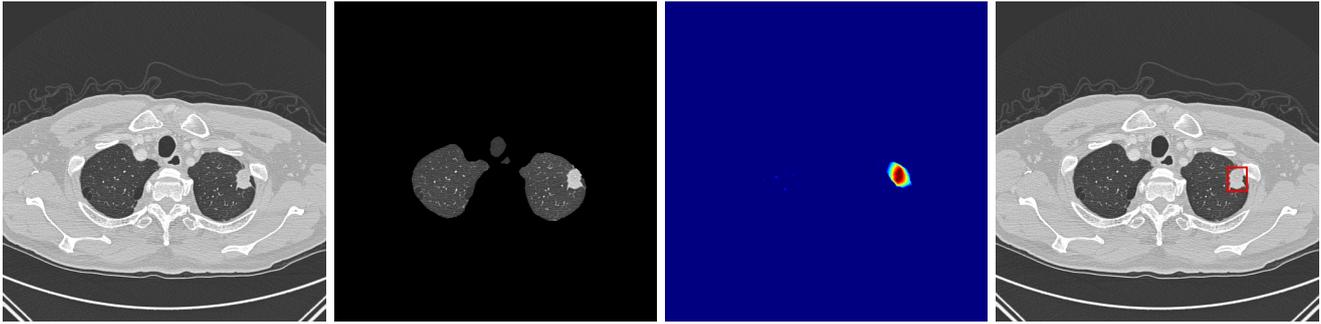


Fig. 1: Illustration of our complete workflow including segmentation, heatmap calculation and classification of the detected tumor. The red marker is the computed indicator for a malign nodule.

results for differentiation between malign and benign nodules (AUC of 92.7%). A segmentation step for a complete tumor recognition system is missing, since the precise tumor position has already been determined by the radiologist. Thus, the work does not provide a system that can automatically segment and classify tumors without the preliminary work of a specialist.

- Zhao et al. [6] present a complete workflow that implements automatic segmentation and separation of tumor tissue using thresholding and morphological operations, without prior knowledge of tumor positions. They achieve an accuracy between 86.8% - 93.9%. They classify tumors of 3 different predefined size groups. This presupposes in turn a manual division of the data, which facilitates the segmentation problem by a possible reduction of parameters. The proposed approach is not fully automated, as for a functioning segmentation information on the position or size of the tumor must be given in advance.

We suggest a system for automated tumor detection, which implements all sub-steps (preprocessing, segmentation, feature extraction and classification) without relying on previous knowledge in terms of tumor type, position or size. Our system is able to automatically detect tumors in lung CT scans and classify them as benign or malignant. In addition to this fully automated approach, we provide a user interface to evaluate results independently, set markers to optimize segmentation results and to select fixed cutouts for classification.

We will evaluate our novel approach using a data set from the SPIE-AAPM Lung CT Challenge [10], [11], [1], which consists of CT scans of 70 patients of different age groups with a slice thickness of 1 mm. For each patient, the scans contain one or more either benign or malignant lung tumors identified by follow-up examinations or pathological assessments by experts. Both the position of the tumor center and the classification into benign or malignant were annotated. The dataset is divided into a calibration dataset containing 10 patients, as well as a test dataset, which covers the remaining 60 patients. The calibration dataset contains CT scans with exactly 5 benign and 5 malignant tumors. The test dataset consists of a total of 73 sections with 36 malignant and 37

benign tumors. The size of the tumors in the dataset varies widely with small tumors being less than 3 mm in diameter and large tumors larger than 35 mm in diameter. The difficulty level of the dataset for the evaluation of benign and malignant tumors can be classified as very demanding. Out of all 11 approaches submitted in this challenge, only 3 achieved an AUC score significantly better than random guessing [10]. The AUC values of radiologist assessments ranged between 0.7 and 0.85. The best participant scored an AUC value of 0.68, see [10].

III. METHODS

This novel approach follows the results of Zayed and Elnemr [4], Han et al. [5], and Zhao et al. [6] and uses texture features for feature extraction. An SVM is used for classification as it proved to be more successful than other classifiers including neural networks, see Bahadure et al. [9]. The tumor detection is based on *multiscale sliding windows*, since this method is independent of the size of the searched object. In a previous segmentation step, the lungs are extracted to reduce the search area. This improves the results and reduces the computing time. For the segmentation the *marker-based watershed transformation*, as supposed by Kulkarni et al. [8], is used. This method can also be used to separate tumors that have grown into the lung wall. The markers are computed in a preliminary step using morphological operations. In addition to the introduced state of the art works, we present a complete system for tumor detection and diagnosis that performs all necessary steps for a tumor recognition system described above and is independent of the tumor size.

Our approach can be divided into three steps:

1) Preprocessing and Segmentation:

In this first step, the lungs are separated from the rest of the tissues and the image background. Using morphological operations and *marker-based watershed transformation* (WST), the image background and the ribcage are removed.

2) Tumor detection:

Using a sliding window approach, texture features are evaluated on different scales to compute heat maps, which are used to identify tumor structures. The heat

maps indicate image areas which contain potentially tumor-like textures without providing information about the malignancy of tumors yet. For later evaluation the user can either use this information to independently mark structures for evaluation or automatically select all interesting image areas for the evaluation based on the prior calculated heat map.

3) Tumor classification:

Benign and malignant nodules can again be differentiated by certain texture features as proposed by [4], [5], [6]. The areas selected by the heat map or manually selected by the user are again evaluated by an SVM using texture features.

The complete workflow is shown in figure 1. The individual steps in the diagnosis process are described individually as such in the following sections.

For the purpose of training and evaluation, we inspected the slices containing the tumor center for all ten patients in the calibration set. We receive one or more grayscale images of 512×512 pixels per patient, depending on how many tumor centers have been annotated. The SVM was trained using the calibration set provided by the challenge. We evaluated our results using the test data set consisting of a total of 73 images. The test set contained 30 images in which tumors were already fused with the lung wall. In the remaining 43 test images, the tumors were isolated inside of the lung. All CT images shown in this paper are either taken from the dataset provided by the challenge [1] or amended by our presented diagnostic system.

Preprocessing and Segmentation

The segmentation of the CT scans isolates the internal lung tissue and facilitates the detection and classification of the tumors. Large areas of the image are removed in this process and do not need to be considered for later computation. The image background followed by the thorax is removed in two steps, using *binarization*, *erosion*, *connected component labeling* and the *marker-based watershed transformation*. These operations were implemented using the OpenCV library (see [12] or <https://opencv.org/>). In addition to a fully automated approach, users are also provided with an interactive mode. Here users can set markers for an automatic separation of tissue based on marker positions which may improve the segmentation results.

CT scans often contain artifacts in the image background. These can be a barrier for further processing and segmentation because they represent separate components that are also recognized as such by the *connected component* algorithm, even though they are not part of the tissue that should be examined. These artifacts are removed from the image background in a first preprocessing step.

In several pictures, the chest adjoins the outer edge of the picture, creating two separate background areas in the upper and lower part of the picture. To prevent this and create a coherent image background, a margin of 5 pixels is set for the outer left and right sides of the image. Image noise is removed using a median filter to improve further processing. From the filtered image, a binary image is calculated using

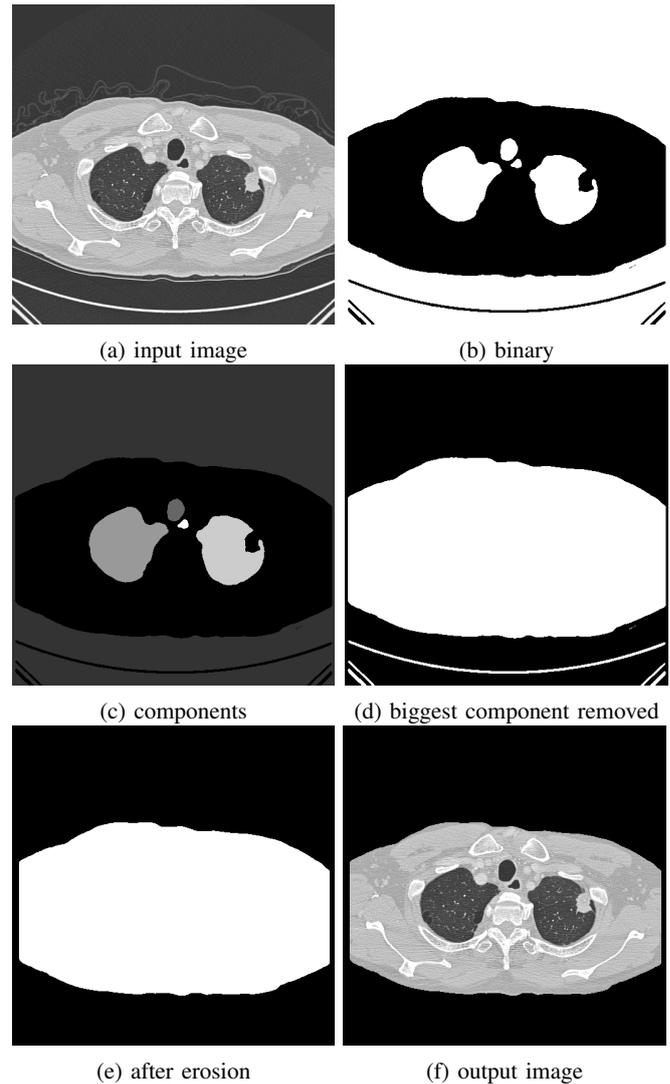


Fig. 2: Removal of the image background: The input image (a) is used to create a binary image (b). Using Connected-component labeling components are determined (c). The biggest component gets removed (d) and the image gets eroded (e). By masking the eroded image on the input we obtain the lung corpus (f).

a threshold intensity value of 130. The *connected component algorithm* is applied to this binary image to identify the largest foreground region as the image background, which is removed in a new binary image. By erosion, all artifacts in the image background can now be removed until only the ribcage is left as a single foreground component. For this purpose, we erode the binary image with an increasingly bigger quadratic kernel starting with a kernel size of 1. After each iteration, the number of foreground components is checked using the *connected component algorithm*. This step is repeated until only a single component, namely the rib cage, is detected. The remaining background is now applied as an image mask to the

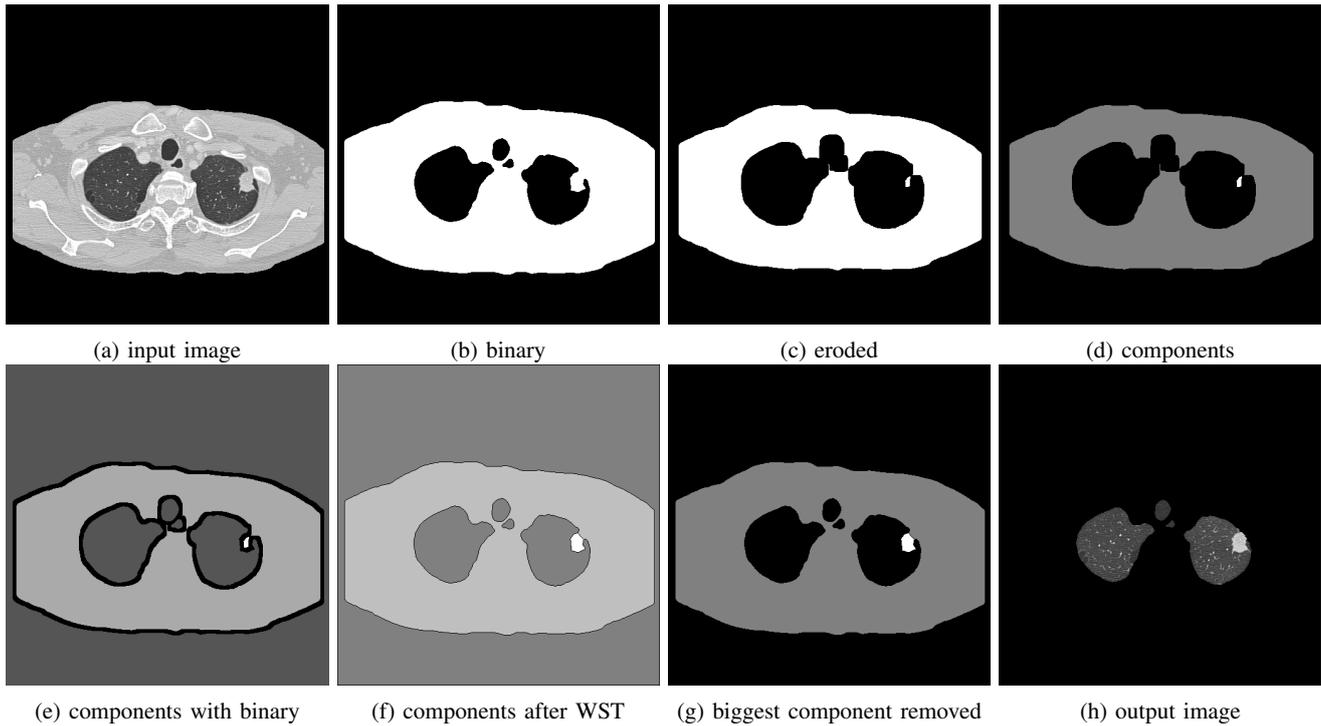


Fig. 3: Removal of the thorax: The input image (a) is used to create a binary image (b). The binary image is now eroded to separate inter-grown tumors from the lung wall (c). By means of Connected-component labeling, all isolated components are determined (d) and combined with the prior calculated binary image (e). After a marker-based watershed transformation (f) the biggest component is removed (g) followed by the second biggest component (thorax). The resulting image only contains the isolated lung tissue (h).

input image. The result is the isolated body scan without the image background. The whole procedure is shown in figure 2.

In the second step, the thorax should be removed without removing any attached tumor structures inside the lung tissue. For this purpose, the output image of the last step is further used as input. The input image is smoothed and binarized as described above. The binary image is then eroded to separate possible tumor structures that are internally connected to the thorax. For the Erosion, a 13×13 pixel kernel is used. Following systematic tests, this kernel size has proven to be optimal in order to successfully separate as many tumor structures of the test data set as possible from the lung wall. Using *connected component labeling*, all isolated structures in the eroded binary image are now identified and saved as markers. These markers are then used in a *marker-based watershed transformation* on the binary image to separate the rib cage from the inner tissue. The different markers spread to all foreground pixels of the previously created binary image. The labeled area with the largest volume is identified as the chest area and is removed from the original image like previously the background. In the finished segmented result image only the two lungs remain. The methodology is illustrated in figure 3.

With the proposed methodology, it is not possible to separate all tumors that are connected to the lung wall. It is necessary that the diameter of the junction between tumor

and lung wall is smaller than the total diameter of the tumor. Otherwise the tumor can not be completely separated from the lung wall by erosion without entirely removing it. As a result, no marker for the WST can be obtained and the tumor assigned to the segment of the thorax after the WST step and is completely removed together with the thorax in the following step. Using the proposed method on the test data 18 out of 30 tumors that were connected to the lung wall can be successfully separated automatically. This corresponds to a loss of 16.44% of all tumors to be segmented in relation to the entire data set of 73 images. If markers are placed manually at critical locations before segmentation 100% of the test images can be successfully segmented.

Tumor Detection

The recognition of tumors on the basis of texture features can be difficult if the size of the tumor is unknown. Considering windows of different sizes, texture features may have different values for the same image coordinate. This is due to the fact that the texture of the window contents changes significantly with different window sizes. Even if a trained classifier gives a positive response to the texture features of the correct window size, the response may be negative if the window is too large or too small.

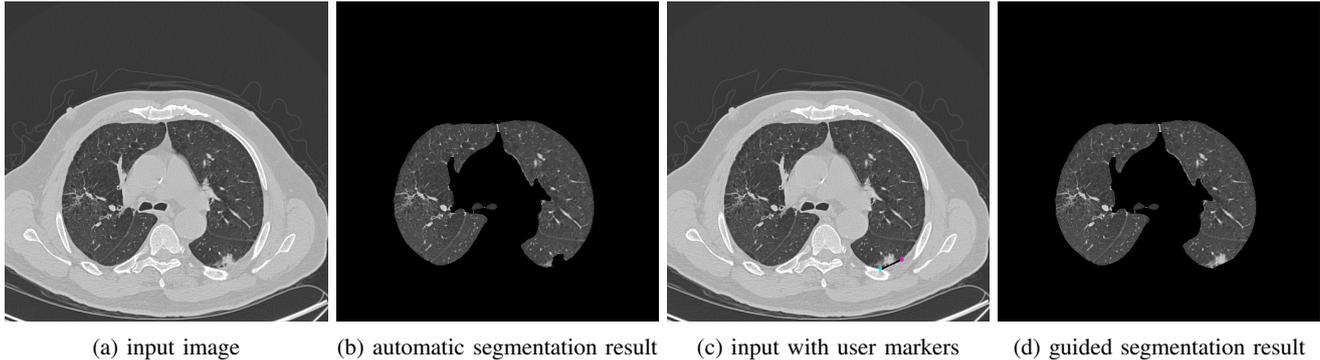


Fig. 4: Guided segmentation using markers provided by the user. The tumor in the input image (a) could not be automatically segmented. However, if markers are provided by the user (c), the segmentation yields a correct result.

To localize the tumor structures, our approach utilizes sliding windows of 11 different scales ranging from 29×29 pixels to 9×9 pixels, which correspond to the maximum and minimum size of all tumors found in the training set. The sliding window iterates through the image from the top left to the bottom right corner. Texture features are extracted from each window. An SVM uses these features to calculate a score and assign the window to a positive or negative class. In a result matrix, the entry corresponding to the central coordinate of the current window is increased if the respective section is assigned to the tumor class. The entry is additionally scaled with the SVM score of the respective window to assign more weight to windows that receive a high rating by the SVM. After generation of the last weighted result matrix, all result matrices are concatenated and the resulting matrix is normalized to a maximum intensity of 255. The resulting matrix is used to create a heat map that identifies image areas with tumor-like texture.

The complexity of the heat map calculation is in $\mathcal{O}(11nm) = \mathcal{O}(n^2)$, where n is the number of image lines, m is the number of image columns, and $n = m$.

For classification purposes, this work uses the SVM-light implementation of Joachims (see [13] or <http://svmlight.joachims.org>). For the training of the SVM, image sections with a size of 9×9 pixels to 29×29 pixels from the 10 images of the calibration dataset were generated for the positive class on the basis of the annotated tumor centers of the dataset per patient for each possible scale. For the negative class, 20 random cutouts of a random size between 9×9 and 29×29 pixels were selected from the rest of the image. Sections for the negative class which contained a tumor center were discarded and regenerated. For all sections, JFeatureLib (see [14] or <https://github.com/locked-fg/JFeatureLib>) was used to extract feature vectors with texture features that were used to train the SVM. For the training of the SVM an RBF kernel with a σ value of 10^{-8} was used. The best kernel and optimal parameters were experimentally determined by optimizing the accuracy on the training data. The accuracy was calculated using leave-one-out cross-validation.

The heat maps generated in the previous step describe

regions, whose texture is most similar to those of tumors. From these regions, excerpts are taken for evaluation in the last step. For this, the minimum bounding box of each isolated area of the heat map is calculated. Boxes that are smaller than 5×5 pixels are discarded because the smallest tumors already have a diameter of at least 9 pixels. Based on the calculated bounding box, the central coordinate of each area is determined. These central coordinates are then used to find the bounding box of the respective component in the input image which corresponds to the area of the heat map. Since the bounding box of the heat map does not always correspond to the full size of the respective components in the input image, this approach has the advantage that the new bounding box fully covers the components in the original image and can thus optimally describe the texture of the respective components.

The various components are already separated by WST and provided with a unique label connected to the prior segmentation step. Based on the previously determined central coordinate, the label and the associated component can be determined. Based on this information, a new bounding box can be calculated which corresponds to the component in the input image. For each window generated, the SVM again calculates a score which is intended to reflect the probability that a tumor is present in the respective window.

IV. RESULTS

The tumor detection was evaluated using 2 different strategies:

Strategy 1: All found windows were treated as tumors. This approach has the advantage of minimizing the number of false negative results. However, as many textures can be recognized as a tumor in some images, the number of false positives also increases significantly.

Strategy 2: Only the window with the highest SVM score is considered. The advantage of this strategy is that as many false positives as possible can be excluded. The disadvantage is that true positives can also be rejected as false negatives. This would be fatal, especially in the case of an actual diseased

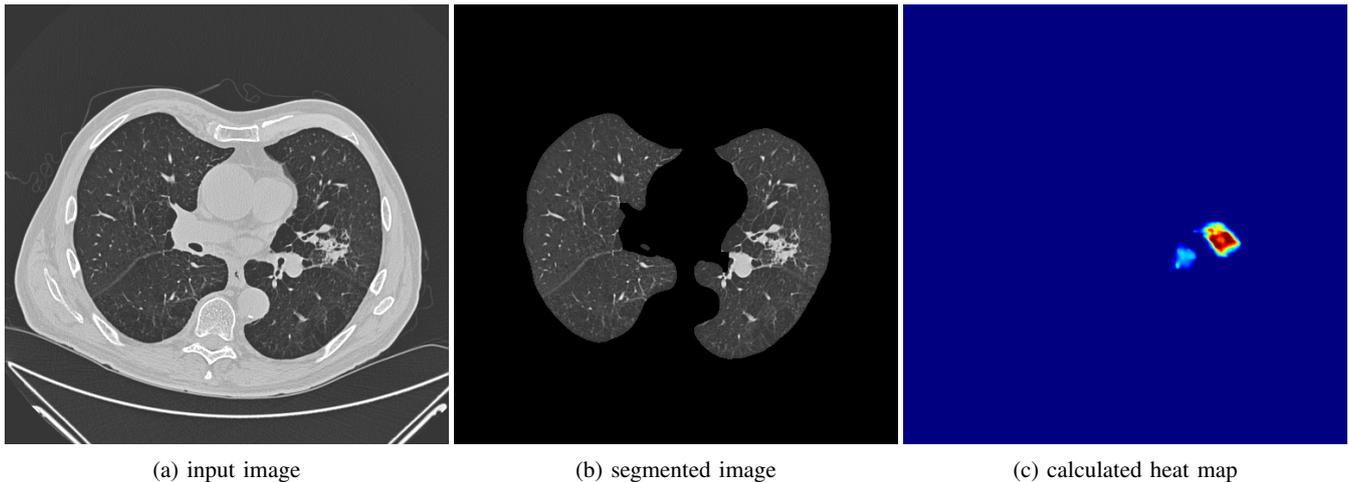


Fig. 5: Segmented image and resulting heat map

patient, since this corresponds to an unrecognized potentially harmful tumor.

An example of a scan in which more than a single tumor is found is shown in figure 6. The bounding boxes of the two identified components are determined by the heat map in the original image and evaluated again by the SVM. In Figure 6, the right calculated bounding box contains a correctly recognized tumor. The left bounding box includes a component that has been erroneously recognized as a tumor. The SVM score of the correctly recognized tumor segment is with 3.12 higher than that of the incorrectly recognized tumor segment with a score of only 0.88.

The results of the window selection are described in table I. For the SPIE-AAPM Lung CT Challenge test dataset, only the tumor centers coordinates are annotated. A window is considered a true positive if it contains the annotated tumor center. All windows that do not contain a tumor center are considered false positives. Tumors that were not detected by a window were considered false negatives. Out of a total of 73 tumors, 59 tumors were detected and 14 tumors were not detected. Of the 59 recognized tumors, 44 had the highest SVM value of any detected windows in each image. Strategy 1 improves the recall by over 20% compared to Strategy 2. However, the precision value is over 40% below the value of Strategy 1. Strategy 2 thus also leads to a higher F-measure. Although Strategy 2 performs statistically better than Strategy 1, it should still be viewed critically for practical application. In the case of an actual application, automatically recognized tumors could once again be confirmed or denied by expert knowledge; a false negative would have far worse consequences in such a scenario, as an unrecognized tumor would in any case be a risk for the patient.

TABLE I: Evaluation of tumor detection

Strategy	TP	FP	FN	Precision	Recall	F-Measure
Strategy 1	59	112	14	34,50%	80,82%	0,4835
Strategy 2	44	15	29	74,58%	60,67%	0,6666

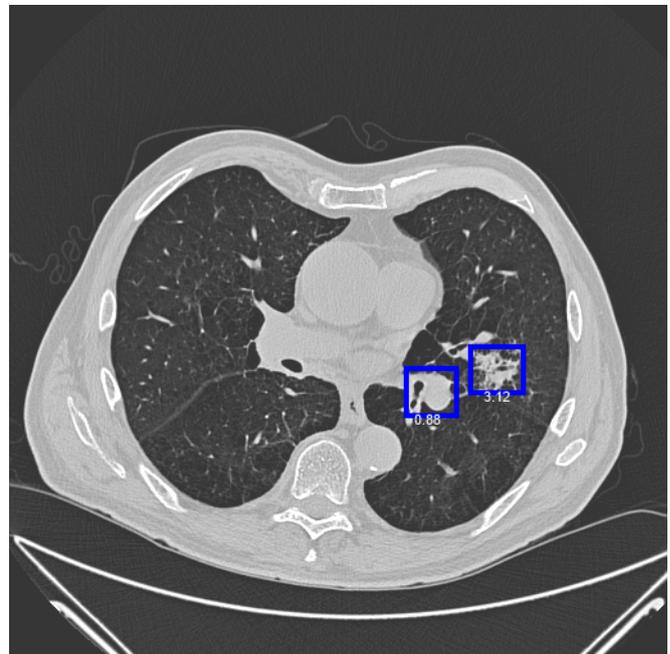


Fig. 6: Two detected tumors with different SVM ratings. The right window with a rating of 3.12 contains a correctly detected tumor, the left window with a rating of 0.88 contains a falsely detected non-tumor structure.

Tumor classification

In addition to the differentiation between tumor and non-tumor tissue, tumor tissue can again be classified as benign and malignant on the basis of the present texture. For this purpose, the image windows previously obtained from the heat map are evaluated by a second SVM. In contrast to the previous step, the textures of the two classes differ only marginally. Feature vectors that have an equal or similar mean in the same dimensions for both classes are difficult

to separate using these features. A reduction of the feature space by removing such dimensions, which are very similar or identical for both classes, can increase the accuracy of the classification. Therefore for the distinction of the two tumor classes only those features are used, whose mean values differ significantly for both classes. Qian Zhao et al. [6] already identified homogeneity, energy, correlation and entropy as the most discriminating features in t-tests in order to distinguish between benign and malignant tumors.

In this work, the mean values of the texture features were analogically compared. By utilizing t-tests, p-values were determined for each feature in order to determine the discriminating characteristics. The features were extracted from windows containing tumors for the 10 patients of the calibration dataset. For the training of the SVM and the classification only features with a p-value of less than 0.05 were used to train the classifier.

TABLE II: Evaluation of p-values for feature selection

Feature	Mean		Variance		p-value
	benign	malign	benign	malign	
Energy	0,014	0,034	0,027	0,026	0,147
Contrast	12,532	9,870	7,082	10,908	0,961
Correlation	250983	27771	189268	119048	0,024
Variance	12004	4797	6579	3526	0,003
Homogeneity	0,366	0,426	0,114	0,105	0,228
Sum Average	29,65	17,60	7,806	5,976	0,001
Sum Variance	118,74	58,28	44,856	86,548	0,223
Sum Entropy	3,032	2,435	0,460	0,533	0,024
Entropy	4,434	3,537	0,724	0,757	0,020
Diff. Variance	5,709	6,790	3,239	6,005	0,632
Diff. Entropy	11,912	11,625	0,270	0,376	0,189
Meas. of Corr. 1	-0,235	-0,198	0,045	0,062	0,238
Meas. of Corr. 2	10,505	10,000	0,289	0,177	0,022
max. Corr. Coeff.	11,007	10,832	0,145	0,200	0,034

In order to evaluate the classification independently of the preliminary step, the tumor windows for evaluation were determined on the basis of the annotated central coordinates. The SVM was trained on the calibration data set based on the previously determined discriminating features. The performance was evaluated on the basis of the 73 sections of the test data set, as intended by the organizers of the challenge. Based on the results of the t-tests presented in Table II, the texture features Correlation, Variance, Average Sum, Sum Entropy, Entropy, Correlation 2, and Maximum Correlation Coefficient were selected for the training of the SVM. The significance of the features correlation, variance and entropy described in [6] can thus be confirmed. It is not possible to confirm the significance of the contrast and energy characteristics which were rejected on the basis of the calculated p-values. One possible explanation for these different outcomes for the two features would be a different methodology for the tumor window selection. Through the evaluation of the entire bounding box, areas of the adjacent background for the classification were considered in this work.

The SVM achieves the best classification results using the RBF kernel with a σ value of 10^{-8} . The kernel with the highest performance and the corresponding optimal parameters were determined experimentally. The test dataset achieved a recall value of 0.75 %, a precision value of 0.5625 %, and an

accuracy of 0.589 %. The ROC calculated for the SVM output has an AUC value of 0.61. The ROC is shown in figure 7.

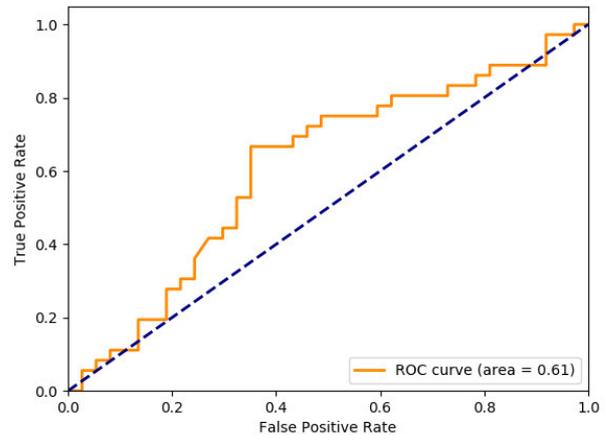


Fig. 7: Calculated ROC curve describing our classification results

Of all 11 methods submitted to the challenge, only a total of 2 achieved an AUC value above 0.61. Compared to the submissions that only used the calibration data set consisting of 10 patients for training, the presented work scores second best. Other submissions used the National Lung Screening Trial (NLST) dataset with 53,454 lung scans of former smokers or the Lung Image Database Consortium (LIDC) dataset with lung CT scans of 1,010 different patients. The submission with the highest achieved AUC value used an unspecified in-house dataset. Most submissions use some form of thresholding or *region growing* for segmentation. The *watershed transformation* used in this work is not used in any of the submitted papers. In addition to our approach, three submissions of the Challenge use an SVM as a classifier. Of all submissions that use an SVM classifier, two achieved a lower AUC value than the presented work; one work achieved the same AUC value. A submission uses a *convolutional neural network* (CNN) trained on the LIDC record as a classifier. However, this work only achieved an AUC of 0.59. The best work achieved an AUC of 0.68 using a *support vector regressor* for classification.

V. DISCUSSION AND FURTHER RESEARCH

In the following section, we will discuss our obtained results and present possible improvements to further enhance our presented methods. We suggest improvements for each individual step which may further increase the accuracy of the presented system.

A. Segmentation

With the presented methodology 83.56% of the test images were successfully segmented. Successful segmentation requires both the separation of lungs and lung wall, as well as the separation of tumors from the lung wall, if they are

interconnected. The biggest challenge of the segmentation has been the separation of lung-walled tumors. While the presented methodology was able to correctly segment all lungs with isolated tumors, 12 out of 30 of the lung wall tumors could only be separated by manually placed markers. In the segmentation step, therefore, it has been shown above all that additional user input can be used to improve the efficiency of the system. For our presented approach the prerequisite for a successful automatic segmentation is a maximum width of the connection region between tumor and lung wall. This width must be less than the diameter of the tumor, since a separation by erosion is otherwise impossible. This problem was solved in this work by an active approach with user input by placing markers at critical junctions. A desirable approach would be able to find these markers fully automatically without user input, whereby the segmentation could also be carried out fully automatically for all special cases. This could be realized by a form of edge tracking, which marks the affected image area in the event of a strong change of the gradient direction.

B. Tumor detection

Using texture features, up to 80.82% of the annotated tumors could be successfully detected and localized, provided that all areas of the heat map were considered for detection. However, this strategy also falsely identifies tumors in many areas of the image. If only the area with the highest SVM score was considered per image, 88.19% of these false positives could be eliminated. However, this strategy reduces the recall to 60.67%. It has thus been shown that an improvement of the recognition accuracy results in a reduction of the recognition rate and vice versa. Future work could build on the results to find methods that eliminate a larger number of false positives without reducing the recall value.

C. Tumor classification

Compared to the other work of the SPIE-AAPM Lung CT Challenge, the proposed methodology has achieved above-average results. The SVM was trained only using the provided calibration set consisting of 10 images. This shows that the presented methodology of classifying texture features by SVM is able to achieve good results even on small training sets. Training with a larger dataset could potentially further improve the classification results.

Currently, only the layer containing the tumor center is used to evaluate the tumor based on its textural features. Fang Han et al. [5] use three-dimensional Haralick features to classify tumors. In their approach, surrounding tissue layers are also considered for the evaluation of the tumor. They increased the AUC value for their dataset to 0.9441 by using three-dimensional Haralick features compared to two-dimensional Haralick features which scored an AUC value of 0.9373. In future work, three-dimensional Haralick features could be utilized to possibly further improve the accuracy of the classification.

VI. CONCLUSION

The results of this work have shown that the presented methodologies can be successfully used to implement a complete system for automatic tumor diagnosis. We received and presented very encouraging results. Texture features can still be considered a strong tool for image classification, even in complex applications like tumor recognition and classification. Furthermore our SVM classifier has proven to be very effective in combination with Haralick features, achieving better results than several other classifiers on the same data set.

REFERENCES

- [1] S. Armato III, L. Hadjiiski, G. Tourassi, K. Drukker, M. Giger, F. Li, G. Redmond, K. Farahani, J. Kirby, and L. Clarke, "Spie-aapm-nci lung nodule classification challenge dataset," *Cancer Imaging Arch*, 2015.
- [2] L. A. Torre, R. L. Siegel, and A. Jemal, "Lung Cancer Statistics," Springer, Cham, 2016, pp. 1–19. [Online]. Available: http://link.springer.com/10.1007/978-3-319-24223-1_{_}1
- [3] Cancer Research UK, "World cancer factsheet," 2012. [Online]. Available: http://www.cancerresearchuk.org/sites/default/files/cs_report_world.pdf
- [4] N. Zayed and H. A. Elnemr, "Statistical Analysis of Haralick Texture Features to Discriminate Lung Abnormalities," *International Journal of Biomedical Imaging*, vol. 2015, pp. 1–7, oct 2015. [Online]. Available: <http://www.hindawi.com/journals/ijbi/2015/267807/>
- [5] F. Han, H. Wang, G. Zhang, H. Han, B. Song, L. Li, W. Moore, H. Lu, H. Zhao, and Z. Liang, "Texture Feature Analysis for Computer-Aided Diagnosis on Pulmonary Nodules," *Journal of Digital Imaging*, vol. 28, no. 1, pp. 99–115, feb 2015. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/25117512http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4305062http://link.springer.com/10.1007/s10278-014-9718-8>
- [6] Q. Zhao, C.-Z. Shi, and L.-P. Luo, "Role of the texture features of images in the diagnosis of solitary pulmonary nodules in different sizes," *Chinese journal of cancer research = Chung-kuo yen cheng yen chiu*, vol. 26, no. 4, pp. 451–8, aug 2014. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/25232219http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4153941>
- [7] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, nov 1973. [Online]. Available: <http://ieeexplore.ieee.org/document/4309314/>
- [8] S. G. Kulkarni and S. B. Bagal, "Techniques for Lung Cancer Nodule Detection: A Survey," *International Research Journal of Engineering and Technology*, pp. 2395–56, 2015. [Online]. Available: <https://irjet.net/archives/V2/i9/IRJET-V2I9323.pdf>
- [9] N. B. Bahadure, A. K. Ray, and H. P. Thethi, "Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM," *International Journal of Biomedical Imaging*, vol. 2017, pp. 1–12, mar 2017. [Online]. Available: <https://www.hindawi.com/journals/ijbi/2017/9749108/>
- [10] S. G. Armato, K. Drukker, F. Li, L. Hadjiiski, G. D. Tourassi, R. M. Engelmann, M. L. Giger, G. Redmond, K. Farahani, J. S. Kirby, and L. P. Clarke, "LUNGx Challenge for computerized lung nodule classification," *Journal of Medical Imaging*, vol. 3, no. 4, p. 044506, dec 2016. [Online]. Available: <http://medicalimaging.spiedigitallibrary.org/article.aspx?doi=10.1117/1.JMI.3.4.044506>
- [11] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, dec 2013. [Online]. Available: <http://link.springer.com/10.1007/s10278-013-9622-7>
- [12] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
- [13] T. Joachims, 2008, <http://svmlight.joachims.org/> Accessed: 2018-02-05.
- [14] <https://github.com/locked-fg/JFeatureLib> Accessed: 2018-02-05.

3rd International Workshop on AI aspects of Reasoning, Information, and Memory

THERE is general realization that computational models of human reasoning can be improved by integration of heterogeneous resources of information, e.g., multidimensional diagrams, images, language, syntax, semantics, memory. While the event targets promotion of integrated computational approaches, we invite contributions from any individual areas related to information, language, memory, reasoning.

TOPICS

We welcome submissions of papers on the following topics, without limiting to them, across approaches, methods, theories, and applications:

- Reasoning systems — theories and applications
- Proof systems and model checkers
- Theories of computation and information
- Interactive computation and reasoning
- Computation and reasoning with heterogeneous information
- Space and time in information, language, memory, and reasoning
- Partiality, underspecification, vagueness, and possibilities
- Detection of and reasoning with inconsistency
- Logic and language — approaches, theories, methods
- Computational morphology, syntax, semantics, and interfaces between these
- Constraint-based and type-theoretic approaches and grammars
- Logical approaches to multilingual processing
- Logical and computational foundations in machine learning and information retrieval
- Mathematics for linguistics and cognitive science
- Reasoning, information, and memory in computational neuroscience and life sciences
- Interdisciplinary approaches to information, language, memory, and reasoning

EVENT CHAIRS

- **Grabowski, Adam**, Institute of Informatics, University of Bialystok, Bialystok, Poland
- **Ishihara, Hajime**, Japan Advanced Institute of Science and Technology, Japan
- **Loukanova, Roussanka**, Stockholm University, Sweden
- **Schwarzeweller, Christoph**, Institute of Informatics, University of Gdansk, Poland

- **van den Herik, Jaap**, Leiden University, The Netherlands

PROGRAM COMMITTEE

- **Akman, Varol**, Ihsan Dogramaci Bilkent University, Turkey
- **Becerra, Leonor**, Jean Monnet University, France
- **Bekki, Daisuke**, Ochanomizu University / JST CREST, Japan
- **Borgefors, Gunilla**, Uppsala University, Sweden
- **Buszkowski, Wojciech**, Adam Mickiewicz University, Poland
- **Cooper, Robin**, University of Gothenburg, Sweden
- **Hellan, Lars**, Norwegian University of Science and Technology, Trondheim, Norway
- **Jiménez López, M. Dolores**, Universitat Rovira i Virgili, Spain
- **Kerber, Manfred**, University of Birmingham, United Kingdom
- **Kornilowicz, Artur**, Institute of Informatics, University of Bialystok, Poland
- **Litak, Tadeusz**, Informatik 8, FAU Erlangen-Nuremberg, Germany
- **Nemoto, Takako**, School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Japan
- **Nilsson, Jørgen Fischer**, Technical University of Denmark, Denmark
- **Penn, Gerald**, University of Toronto, Canada
- **Retoré, Christian**, Université de Montpellier & LIRMM-CNRS, France
- **Rocha, Ana Paula**, University of Porto, LIACC / FEUP, Portugal
- **Sailer, Manfred**, Goethe-Universität Frankfurt am Main, Germany
- **Salvati, Sylvain**, Laboratoire Bordelais de Recherche en Informatique, Unité Mixte de Recherche CNRS (UMR 5800), France
- **Schwichtenberg, Helmut**, Mathematisches Institut der Universität München, Germany
- **Villadsen, Jørgen**, Technical University of Denmark, Denmark

Mizar Set Comprehension in Isabelle Framework

Karol Pąk

University of Białystok,

Ciołkowskiego 1M, 15-245 Białystok, Poland

Email: pakkarol@uwb.edu.pl

Abstract—The Mizar project from its beginning aimed to make a highly human oriented proof environment where the proof style closely reflects the informal proofs style. The support is reflected in the size of the largest consistent formal library—Mizar Mathematical Library (MML). However, the Mizar system is the only tool that provides full verification and further development of the MML. In this paper, we present the progress in the development of the Isabelle/Mizar project whose main goal is independent cross-verification of the MML in Isabelle. We focus on Mizar set comprehension operators that allow defining sets that satisfy a given predicate. The development already covers simple cases where the arity of predicates is limited to two. We propose an infrastructure that provides a more elegant and recursive approach to construct and to provide the main property of set comprehension operators.

I. INTRODUCTION

Mizar Mathematical Library (MML) [1] is one of the most recognizable features of the Mizar system. Developed for almost three decades the library contains today more than 1300 articles, 60000 proved theorems and covers many areas of today’s mathematics from algebra, analysis, topology including topological manifolds [2] and lattice theory [3] that have not been formalized elsewhere. Therefore, it is not a surprise that there exists a number of external tools that explore the content of the MML to ensure human-readable access, starting with automatically generated articles in the Journal of Formalized Mathematics, searching tools as MML Query [4], variants of XML format [5] and MMT logical framework [6].

On the other hand, the MML is often used as an extensive theorems database, for instance, in the process of comparing the performance of leading systems of Automatic Theorem Proving (ATP) as well as a training data in machine learning, especially for developing and testing premise selection methods [7]. However, the Mizar logic is a serious problem for today’s efficient first-order ATP systems. It is important to note that the Mizar is essentially a first-order system that is based on the set theory, but the Mizar logic goes a little bit beyond first-order in two cases:

- the Mizar schemes that are second-order theorems parameterized by the predicates and functions,
- the Mizar set comprehensions (referred to as *Fraenkel* in the Mizar literature [8]) that allow defining sets of terms whose arguments have given types and satisfy a given predicate.

The paper has been supported by the resources of the Polish National Science Center granted by decision nr DEC-2015/19/D/ST6/01473.

Therefore, to translate and further to cross-verify the content of the MML we have to choose between first and higher order logic. Obviously, first-order logic is welcome from the ATP point of view, but currently existing translations omit each problem where second-order constructions occur or they need to be expressed in first-order logic with the support of a potentially infinite number of axioms [9]. On the other hand, second-order Mizar problems have been cross-verified by C. Brown [10] using higher-order automated theorem provers Satallax and LEO-II with the support of only a few additional axioms.

Isabelle/Mizar is a project whose main goal is an automatic translation of the Mizar proof scripts from the MML to the Isabelle framework, enabling cross-verification of the obtained scripts, but in contrast to the existing translations it tries to preserve types, commands and the structure of proofs originally used [11], [12]. The project is also a unique from the point of view of the order of logic. Namely, our object logic created in Isabelle that expresses that the foundations of the Mizar logic can be both an extension of first-order and higher-order logic, that is, a user can switch between the dependency on relatively poor Isabelle/FOL and the most developed Isabelle object logic Isabelle/HOL [13].

In this paper, we discuss the progress in the Isabelle/Mizar project in relation to the development of set comprehensions. In our previous work [14] we proposed an equivalent of these sets that can be defined as a meta-functor independently for every arity of relevant predicates. Unfortunately, proofs of such n -arity functor correctness require a lot of effort especially in the case of predicates with many arguments. We will, therefore, propose an infrastructure for a more elegant *recursive* proof of correctness that is able to apply the proven property of n -ary meta-functor to justify corresponding property of $(n+1)$ -ary one. We investigate the efficiency of our procedure up to the maximum arity of the set comprehension used in the MML. Currently, the maximum required n is 6.

In Section II we discuss existing methods that try to express more advanced Mizar concepts in first-order and higher-order systems. We mainly focus on solutions used to express the Mizar set comprehension operators and the number of additional axioms introduced for this purpose. After a short introduction of the axiomatization used in our Isabelle/Mizar project in Section III, we describe our concept of the Mizar set comprehension in Section IV. The particular contributions of this paper are:

- We propose a concept of the product of Mizar types that

is expressed in our semantics that is slightly more liberal than the Mizar one. We use the concept in a new approach to define Mizar set comprehension in a clear and elegant way.

- We investigate the possibilities of our approach to prove recursively the main property of the Mizar set comprehension operators, i.e., every set comprehension determined by given functor, universe and predicate can be replaced by a new constant whose members are exactly the values of the function at each element of the universe that satisfies the predicate.

II. SOLUTIONS IN EXISTING MIZAR TRANSLATIONS

A lot of work has been done to explore the MML by external tools that struggle with many Mizar problems. J. Urban [15] created the largest and the most comprehensive export of MML, initially to the TPTP untyped first-order language where each higher-order problems related to the set comprehension and schemes have been omitted. To cover omitted cases he uses the standard set-theoretic elimination procedure and introduces a dedicated extension of the TPTP language to make the entire MML available for first-order ATPs as a part of the Mizar Problems for Theorem Proving (MPTP) project [9]. Theoretically, all second-order problems could be completely removed from the representation of the MML using the following two rules:

- every reference to a given scheme can be redirected to a copy of the scheme where the occurring second-order variables have been instantiated by the corresponding predicates and functions determined in the context of the reference,
- every set comprehension can be replaced by a new constant with an appropriate property that is guaranteed by the Replacement axiom of Tarski-Grothendieck set theory.

Obviously, the first solution generates a very large expansion, since schemes in most cases refer to other schemes in their justification. Additionally, the Replacement axiom that is originally formulated as a scheme in the MML

```
scheme :: TARSKI_0 : sch 1
Replacement {A () → set, P[object, object]} :
  ex X being set st for x being object holds
    x in X iff ex y being object st y in A () & P[y, x]
provided
  for x, y, z being object st P[x, y] & P[x, z]
  holds y = z;
```

has to be replaced by a potentially infinite number of instances of the axiom. These are necessary to decode the information. The expression $A () \rightarrow \text{set}$ declares a “second-order” 0-arity functor that, in this case, trivializes to a constant and can be instantiated by a term of the type `set`; and the expression $P[\text{object}, \text{object}]$ that declares a “second-order” 2-arity predicate that semantically can be instantiated by a formula with two free variables of the type `object`. The second rule also generates a potentially infinite number of axioms, since

the property of the new constant that replaces a given set comprehension can be introduced as an axiom or proven using the Replacement axiom.

A different approach to solve second-order Mizar problems has been proposed by Kunčar [16] who tried to express the content of the MML in the type system of HOL Light. Obviously, the set comprehension operators and schemes can be naturally expressed in higher-order logic. However, the approach proposed by Kunčar was not able to cover more advanced features of the Mizar type system and finally was only sufficient to translate the first few simpler theories. A successful attempt to cover second-order Mizar problems has been done by C. Brown and J. Urban [10] where second-order Mizar problems have been cross-verified using higher-order automated theorem provers Satallax and LEO-II. However, even in this case the set comprehension operators have been axiomatized instead of defined, using a family of constants repSep_n that correspond to the n -arity set comprehension operators.

III. MIZAR FOUNDATIONS IN ISABELLE

In our previous work [14], we defined a unique equivalent of the Mizar foundations as an object logic in the Isabelle logical framework that includes several definitional mechanisms, the Mizar dependent type system including the structure types as well as the second-order concepts. This equivalent is a result of many experiments whose main goal was to simultaneously express each Mizar components and minimize the number of additional axioms and constants.

The current version of our semantic model of Mizar based on the following Isabelle meta-level types and meta-level constants:

```
typedecl Set
typedecl Ty
consts
  ty_membership :: Set ⇒ Ty ⇒ o           (infix be 90)
  define_ty :: Ty ⇒ (Set ⇒ o) ⇒ (Set ⇒ o) ⇒ Ty
  choice :: Ty ⇒ Set                       (the _)
```

where `Set` corresponds to Mizar terms, `Ty` corresponds to Mizar types, `ty_membership` specifies the relation between terms and types, `define_ty` allows to define types, and `choice` is the choice operator. Note that Mizar distinguishes syntactically types for two kinds: modes that require the existence and adjectives that can restrict modes. We have provided this division in our logical framework before [17], but we have combined these types to simplify our model. To preserve the Mizar semantics we define a meta-predicate

$\text{inhabited}(D) \longleftrightarrow (\exists_M x. x \text{ be } D)$

and assume it defining the bounded quantifiers

```
inhabited(D) ⇒
  Ball(D, P) ⇔ (∀_M x. x be D → P(x))
inhabited(D) ⇒
  Bex(D, P) ⇔ (∃_M x. x be D ∧ P(x))
```

where \forall_M, \exists_M correspond to the standard universal and existential quantifiers of the logic (either Isabelle/FOL or Isabelle/HOL), respectively.

Then to specify all necessary dependencies between terms and types as well as the standard axiom of choice we introduce *only* two axioms that extend the MML axioms, that is, are defined in three axiomatic Mizar articles and are `HIDDEN`, `TARSKI_0`, and `TARSKI_A`, are sufficient to introduce a full semantic model of Mizar. It is important to note that keeping such a small number of axioms is one of the main goals of our project.

axiomatization where

```
def_ty_property: T ≡ define_ty(parent, cond, property) ⇒
  (x be T → x be parent ∧ (cond(x) → property(x))) ∧
  (x be parent ∧ cond(x) ∧ property(x) → x be T) ∧
  (x be parent ∧ ¬cond(x) → inhabited(T)) and
choice_ax: inhabited(M) ⇒ (the M) be M
```

Note that the `def_ty_property` axiom seems to be unnecessarily complicated and could be replaced by a stronger formula $T \equiv \text{define_ty}(\text{property}) \Rightarrow x \text{ be } T \iff \text{property}(x)$. However, our experience has shown that our formulation is weaker but sufficient to define all the necessary concepts. For example, we use the `def_ty_property` axiom to define the negation of type, the intersection of types but also in the case of more advanced concepts, for instance, the conditional functor definitions where meaning (`prop`) of defined functor (`df`) is formulated under some assumption (`as`).

definition NON (non -)

```
where non A ≡ define_ty(object, λ_. True, λ x . ¬ x is A)
```

definition ty_intersection (infixl | 100) **where**

```
t1 | t2 ≡ define_ty(object, λ_. True, λ x. x be t1 ∧ x be t2)
```

abbreviation func_assume_means_prefix

```
(assume _ func _ → _ means _ [0,0,0,0] 10)
```

```
where assume as func df → ty means prop ≡
df = the define_ty(ty, λ_. as, prop)
```

It is also important to note that in our approach we use the MML axioms or even the first few re-formalized articles of the MML to define as well as to provide properties of selected concepts, for instance, we use the root of the Mizar type (`object`) in the above definitions.

IV. MIZAR SET COMPREHENSIONS IN ISABELLE

As it has been shown in Section II the Mizar set comprehension is one of the two second-order Mizar concepts that require a lot of effort in any attempt to cross-verify the MML.

Generally, it allows to use a set of terms $F(v_1, \dots, v_n)$ whose arguments have given types ($v_i \text{ be } \Theta_i$ for $i = 1, 2, \dots, n$) and satisfies the formula $P[v_1, \dots, v_n]$. Note that the Mizar semantic does not allow to define this operator directly in a Mizar script (for more detail see [18]). Therefore, the operator is built-in and is automatically expanded in terms of set membership as follows:

```
x in {F(v1, ..., vn) where v1 is Θ1, ..., vn is Θn:P[v1, ..., vn]}
```

iff

```
ex v1 be Θ1, ..., vn be Θn st x = F(v1, ..., vn) & P[v1, ..., vn]
```

Obviously such a set is guaranteed to exist by the Replacement axiom but only if every type Θ_i has sethood property to avoid Russell's paradox.

definition sethood_prop **where**

```
sethood_prop(M) ≡ ∃ X:set. ∀ x: M. x in X
```

For example, if a type Θ has sethood property, then the existence of the set $\{F(v) \text{ where } v \text{ is } \Theta : P[v]\}$ is a direct consequence of the Replacement axiom substituted by the set of all objects of the type Θ and the predicate $\lambda x y. x = F(y) \ \& \ P[y]$. However, the construction of the suitable set is generally a laborious process, since we need to construct the Cartesian product of sets that cover particular types directly from axioms. By using our re-formalization of the MML in the Isabelle/Mizar system we can reduce the size of such a justification using directly the Cartesian product defined originally in the Mizar script `ZFMISC_1` but the justification is still quite tedious.

A. Recursive Justification of Freankel Obligations

A naive approach to constructing $(n+1)$ -ary set comprehension operators using n -ary one fails in the original Mizar semantics since we cannot define there the product types. However, our semantics is slightly more liberal than that of Mizar and it can be done using the `def_ty_property` axiom as follows

definition ProdType_prefix (- × -)

```
where A × B ≡
```

```
define_ty(object, λ_. True, λ x. x be pair ∧ x'1 be A ∧ x'2 be B)
```

where the pair type corresponds to the Mizar attribute pair and $x'1, x'2$ correspond to the left and right projection of a given term x that can be represented as a pair. Note that the attribute and projections are originally defined in the Mizar article `XTUPLE_0`. We give as an example of our re-formulation definitions of the pair and the left projection.

mdef xtuple_0.def.1 (pair) **where**

```
attr pair for object means
```

```
(λ X. ex x1,x2 be object st X=[x1,x2])
```

mdef xtuple_0.def.2 (- '1) **where**

```
mlet x be object
```

```
assume x is pair func x '1 → object means
```

```
(λ it. for y1,y2 be object st x=[y1,y2] holds it = y1)
```

Then, based on the selected re-formalized theorems including properties of the Cartesian product we provide that the product of inhabited types is inhabited as well as that the product of types that have sethood property also has the property

lemma PT_inhabited:

```
assumes inhabited(A) inhabited(B)
```

```
shows inhabited(A×B)
```

lemma PT_sethood:

assumes inhabited(A) inhabited(B)
 sethood_prop(A) sethood_prop(B)
shows sethood_prop(A×B)

Next, we prove the PT_rule lemma that specifies a dependence between a formula where an existential quantifier binds a variable of a product type $\Theta_1 \times \Theta_2$ and a corresponding formula where two quantifiers have been used to bind separately variables of types Θ_1, Θ_2

lemma PT_rule:

assumes inhabited(T1) inhabited(T2)
shows $(\exists x:T1 \times T2. \text{uncurry}(P)(x)) \longleftrightarrow$
 $(\exists x1:T1. \exists x2:T2. P(x1,x2))$

where the uncurry operator is defined as follows:

abbreviation

$\text{uncurry}(P) \equiv \lambda x. P(x'1,x'2)$

The PT_rule lemma can now be practically used to provide a basic property of 2-arity set comprehension operator based on the corresponding property of 1-arity ones.

theorem Fraenkel2E:

assumes inhabited(T1) inhabited(T2)
 sethood_prop(T1) sethood_prop(T2)
shows x in Fraenkel1($\text{uncurry}(F), T1 \times T2, \text{uncurry}(P)$)
 $\longleftrightarrow (\exists y1:T1. \exists y2:T2. x = F(y1,y2) \wedge P(y1,y2))$
by (rule lfft[OF _ Fraenkel1], rule lfft[OF _ PT_rule],
 auto simp add: assms PT_sethood PT_inhabited)

It is important to note the justification is a single Isabelle tactic where we use all lemmas formulated above. Additionally, modifying only the reference to the theorem we can easily cover all cases up to the arity equals 6 (for more detail see our formalization). Then it is easy to see that every set comprehension operator can be defined as abbreviations for an appropriately substituted 1-arity operator

$\text{Fraenkel}_n(F, \Theta_1, \Theta_2, \dots, \Theta_n, Q) \equiv$
 $\text{Fraenkel1}(\underbrace{\text{uncurry}(\dots(\text{uncurry}(F))\dots)}_{n-1\text{-times}},$
 $\Theta_1 \times \Theta_2 \times \dots \times \Theta_n,$
 $\underbrace{\text{uncurry}(\dots(\text{uncurry}(Q))\dots)}_{n-1\text{-times}})$

V. CONCLUSION

We have presented the progress in our project aiming to cross-verify the MML in Isabelle. In relation to our previous work [14] the proposed recursive approach is an important step forward in defining more clearly Mizar set comprehension operators. We have improved our solution in two aspects. Namely, we indicate how to define more advanced cases based only on the simplest case, i.e., the 1-arity set comprehension operator and we reduce the proof of the main property of Mizar set comprehension in more advanced cases to a single Isabelle tactic. The detail of our formalization is available at:

<http://alioth.uwb.edu.pl/~pakkarol/fedcsis2018/>

REFERENCES

- [1] G. Bancerek, C. Byliński, A. Grabowski, A. Kornilowicz, R. Matuszewski, A. Naumowicz, and K. Pąk, "The role of the Mizar Mathematical Library for interactive proof development in Mizar," *Journal of Automated Reasoning*, 2017. doi: 10.1007/s10817-017-9440-6. [Online]. Available: <https://doi.org/10.1007/s10817-017-9440-6>
- [2] K. Pąk, "Topological manifolds," *Formalized Mathematics*, vol. 22, no. 2, pp. 179–186, 2014. doi: 10.2478/forma-2014-0019
- [3] G. Bancerek and P. Rudnicki, "A Compendium of Continuous Lattices in MIZAR," *J. Autom. Reasoning*, vol. 29, no. 3–4, pp. 189–224, 2002.
- [4] —, "Information retrieval in MML," in *Mathematical Knowledge Management, MKM 2003*, ser. LNCS, A. Asperti, B. Buchberger, and J. H. Davenport, Eds., vol. 2594. Springer, 2003. doi: 10.1007/3-540-36469-2_10. ISBN 3-540-00568-4 pp. 119–132. [Online]. Available: https://doi.org/10.1007/3-540-36469-2_10
- [5] J. Urban, "XML-izing Mizar: Making semantic processing and presentation of MML easy," in *Mathematical Knowledge Management (MKM 2005)*, ser. LNCS, M. Kohlhase, Ed., vol. 3863. Springer, 2005. ISBN 3-540-31430-X pp. 346–360.
- [6] M. Iancu, M. Kohlhase, F. Rabe, and J. Urban, "The Mizar Mathematical Library in OMDoc: Translation and applications," *J. Autom. Reasoning*, vol. 50, no. 2, pp. 191–202, 2013. doi: 10.1007/s10817-012-9271-4
- [7] C. Kaliszzyk and J. Urban, "Mizar 40 for Mizar 40," *J. Autom. Reasoning*, vol. 55, no. 3, pp. 245–256, 2015. doi: 10.1007/s10817-015-9330-8
- [8] A. Grabowski, A. Kornilowicz, and A. Naumowicz, "Four decades of Mizar," *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 191–198, 2015. doi: 10.1007/s10817-015-9345-1
- [9] J. Urban and G. Sutcliffe, "ATP-based cross-verification of Mizar proofs: Method, systems, and first experiments," *Math. in Computer Science*, vol. 2, no. 2, pp. 231–251, 2008. doi: 10.1007/s11786-008-0053-7
- [10] C. E. Brown and J. Urban, "Extracting higher-order goals from the Mizar Mathematical Library," in *Intelligent Computer Mathematics (CICM 2016)*, ser. LNCS, M. Kohlhase, M. Johansson, B. R. Miller, L. de Moura, and F. W. Tompa, Eds., vol. 9791. Springer, 2016. doi: 10.1007/978-3-319-42547-4_8 pp. 99–114.
- [11] C. Kaliszzyk, K. Pąk, and J. Urban, "Towards a Mizar environment for Isabelle: Foundations and language," in *Proc. 5th Conference on Certified Programs and Proofs (CPP 2016)*, J. Avigad and A. Chlipala, Eds. ACM, 2016. doi: 10.1145/2854065.2854070 pp. 58–65.
- [12] C. Kaliszzyk and K. Pąk, "Progress in the independent certification of Mizar Mathematical Library in Isabelle," in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2017. doi: 10.15439/2017F289 pp. 227–236.
- [13] L. C. Paulson and J. C. Blanchette, "Three years of experience with Sledgehammer, a Practical Link Between Automatic and Interactive Theorem Provers," in *Workshop on the Implementation of Logics, IWIL 2010*, ser. EPIc Series, G. Sutcliffe, S. Schulz, and E. Ternovska, Eds., vol. 2. EasyChair, 2010, pp. 1–11.
- [14] C. Kaliszzyk and K. Pąk, "Isabelle formalization of set theoretic structures and set comprehensions," in *Mathematical Aspects of Computer and Information Sciences, MACIS 2017*, ser. LNCS, J. Blamer, T. Kutsia, and D. Simos, Eds., vol. 10693. Springer, 2017, pp. 163–178.
- [15] J. Urban, "MPTP - motivation, implementation, first experiments," *J. Autom. Reasoning*, vol. 33, no. 3-4, pp. 319–339, 2004. doi: 10.1007/s10817-004-6245-1. [Online]. Available: <https://doi.org/10.1007/s10817-004-6245-1>
- [16] O. Kunčar, "Reconstruction of the Mizar type system in the HOL Light system," in *WDS Proceedings of Contributed Papers: Part I – Mathematics and Computer Sciences*, J. Pavlu and J. Safrankova, Eds. Matfyzpress, 2010, pp. 7–12.
- [17] C. Kaliszzyk and K. Pąk, "Presentation and manipulation of Mizar properties in an Isabelle object logic," in *Intelligent Computer Mathematics - CICM 2017*, ser. LNCS, H. Geuvers, M. England, O. Hasan, F. Rabe, and O. Teschke, Eds., vol. 10383. Springer, 2017. doi: 10.1007/978-3-319-62075-6_14 pp. 193–207.
- [18] A. Grabowski, A. Kornilowicz, and A. Naumowicz, "Mizar in a nutshell," *J. Formalized Reasoning*, vol. 3, no. 2, pp. 153–245, 2010. doi: 10.6092/issn.1972-5787/1980

Inference rules for OWL-P in N3Logic

Dominik Tomaszuk

Institute of Informatics, University of Białystok
ul. Ciołkowskiego 1M, 15-245 Białystok, Poland
Email: d.tomaszuk@uwb.edu.pl

Abstract—This paper presents OWL-P that is a lightweight formalism of OWL2. Before proposing our solution we have analyzed the OWL fragment that is actually used on the Web. OWL-P supports easy inferences by omitting complex language constructs. Moreover, we present inference rules for the proposal. Our formalization is based on Notation 3 Logic, which extended RDF by logical symbols and created the Semantic Web logic for deductive RDF graph stores. We also tested experimentally our OWL-P how it deals with real data for reasoning.

I. INTRODUCTION

RESOURCE Description Framework (RDF) is a general method for conceptual description or modeling of information that is implemented in web resources. RDF Schema (RDFS) extends RDF to classes providing basic elements for the description of vocabularies. OWL adds more vocabulary for describing properties and classes i.e. relations between classes, cardinality, and richer typing of properties. Unfortunately, OWL has high worst-case complexity results for key inference problems. The complexity of a fully compliant implementation is considered high [2]. The largest part of this are blank nodes and lists. To overcome this problem we propose a lightweight OWL 2.0 profile called OWL-P.

A rule is perhaps one of the most understandable notions in computer science. It consists of a condition and a conclusion. If a condition that is checkable in some dataset holds, then the conclusion is processed. RDF(S) and OWL entailments can work in the same way.

The paper is constructed according to sections. In Section II we present RDF and Notation 3 Logic concepts. Section III is devoted to related work. In Section IV we present empirical study about OWL elements, OWL-P in the context of OWL2 profile, and comparison to other profiles. In this Section we discuss support of our proposal in existing RDF graph stores. Section V presents reasoning experiments of our OWL profiles. The paper ends with conclusions.

II. RDF AND NOTATION3

The RDF data model rests on the concept of creating web-resource statements in the form of subject-predicate-object expressions, which in the RDF terminology, are referred to as *triples* (or *statements*).

An RDF triple comprises a subject, a predicate, and an object. In [3], the meaning of subject, predicate and object is explained. The *subject* denotes a resource, the *object* fills the value of the relation, the *predicate* refers to the resource's

characteristics or aspects and expresses a subject – object relationship. The predicate denotes a binary relation, also known as a property. More details are presented in [4].

On the other hand, in the Semantic Web environment there is a Notation3 format, which offers a human-readable serialization of RDF model and it also extended RDF by logical symbols and created a new Semantic Web logic called Notation3 Logic (N3Logic). Following [5], we provide definitions of N3Logic below.

Definition 1 (N3Logic alphabet): A N3Logic alphabet A_{N3} consists of the following disjoint classes of symbols:

- 1) a set \mathcal{I} of Internationalized Resource Identifier (IRI) symbols beginning with < and ending with > ,
- 2) a set \mathcal{L} of literals beginning and ending with " ,
- 3) a set \mathcal{V} of variables, $\mathcal{V} = \mathcal{B} \cup V_U$, where \mathcal{B} is a set of existential variables (blank nodes in RDF-sense) start with $_$: and V_U is a set of universal variables start with ? ,
- 4) brackets { , } ,
- 5) a logical implication => ,
- 6) a period . ,
- 7) a keyword @false .

Remark 1: Notation3 allows to abbreviate IRIs by using prefixes. Instead of writing <http://example.com>, we can write ex: .

Definition 2 (Expression): Each IRI, variable and literal is an expression.

Definition 3 (Formula): $\{f\}$ is an expression called formula.

Definition 4 (Implication): $f_1 \Rightarrow f_2$ is a formula called implication.

In Notation3 all expressions can be in all positions of atomic formulas i.e. IRIs, literals, and variables can be subjects, objects or predicates.

Definition 5 (Interpretation): Let V be the vocabulary. An Interpretation V is $I = \langle R^I, E^I, I^I \rangle$, where:

- 1) R^I is a (nonempty) set of resources (the universe of I),
- 2) E^I is a predicate function, $E^I : R^I \rightarrow 2^{R^I \times R^I}$,
- 3) I^I is a interpretation function, $I^I : V \rightarrow R^I$.

We define a simple Notation3 semantics bellow, which is simplified definition of Notation3 semantics [5] that do not support quantification.

Definition 6 (Simple Notation3 semantics): Let I be an interpretation of A_{N3} and f be a formula. Then it satisfies the following conditions:

This paper is an extended version of a paper published in [1].

- 1) If f is $s_{N3} p_{N3} o_{N3}$, then $I \models s_{N3} p_{N3} o_{N3}$ iff $(I^I(s_{N3}), I^I(o_{N3})) \in E^I(I^I(p_{N3}))$,
- 2) If f is $\{f_1\} \Rightarrow \{f_2\}$, then $I \models \{f_1\} \Rightarrow \{f_2\}$ iff $I \models f_2$ if $I \models f_1$.

Number 1 of the definition respects the atomic formulas, which are triples consisting of subject, predicate and object. They can be intuitively seen as first order formulas like *predicate(subject, object)*. Number 2 of the definition respects the implications.

III. RELATED WORK

Apart from Notation3, there are other rule-based inference engines formats for the Semantic Web, such as: FOL-RuleML [6], SWRL [7], RIF [8], [9], R-DEVICE [10], TRIPLE [11], Jena rule¹ and SPIN [12].

FOL-RuleML (First-order Logic Rule Markup Language) [6] is a rule language for expressing first-order logic for the web. It is a sublanguage of RuleML [13]. In FOL-RuleML each of the rules consists of a set of statements called an atom. The atom is a form that consists of objects, which are individuals or variables, and the relation between them.

SWRL (Semantic Web Rule Language) [7] is based on OWL [14] and Datalog RuleML, which is a sublanguage of the RuleML. Moreover, RuleML contents can be parts of SWRL content. Both in RuleML and SWRL logical operators and quantifications are supported. SWRL extends the set of OWL axioms to include Horn-like rules. SWRL axioms consist of OWL, RDF or rules. A relation can be an IRI, a data range, an OWL property or a built-in relation. An object can be a variable, an individual, a literal value or a blank node.

RIF (Rule Interchange Format) [8], [9] is a standard for exchanging rules among disparate systems. It focuses on exchange rather than developing a single one-fits-all rule language. It can be separated into a number of parts, RIF-core [15] which is the common core of all RIF dialects, RIF-BLD (Basic Logic Dialect) [16] comprising basic dialects (i.e. Horn rules) for writing rules, RIF-PRD [17] (Production Rule Dialect) for representing production rules and RIF-DTB (Datatypes and Built-in Functions) [18] comprising a set of datatypes and built-in functions.

R-DEVICE [10] is a deductive rule language for reasoning about RDF data. In R-DEVICE RDF predicates are accomplished as slots with multiple values and resources are represented as the values of RDT types. It supports a second-order syntax, where variables can range over classes and properties. It uses a RuleML-like syntax.

TRIPLE [11] is an RDF rule (query, inference, and transformation) language, with a layered and modular nature. It is based on Horn Logic [19] and F-Logic [20]. Rules in TRIPLE are used for transient querying and cannot be used for defining and maintaining views.

SPIN (SPARQL Inferencing Notation) [12] is a constraint and SPARQL-based rule language for RDF. It can link class with queries to capture constraints and rules which describe the

behavior of those classes. SPIN is also a method to represent queries as templates. It can represent SPARQL statement as RDF triples. That proposal allows to declare new SPARQL functions.

Jena rule is a rule format used only by inference engine in the Jena framework [21]. It uses an RDF-like syntax. It uses triple statements. It is similar to Notation3 Logic but in Jena rule a name of the rule can be defined in a rule. There are not any formula notations. Moreover, built-in functions can be written in function terms. More details are presented in [22].

On the other hand, there are several OWL profiles: RDFS++ [23], L2 [24], RDF 3.0/OWLPrime [25], OWLSIF/pD* [26], OWL LD [27] and OWL-RL [28]. RDFS++ and L2 support basic terms. The first one is devoted to AllegroGraph² and the second is thought to have the greatest possible support. More advanced are RDF 3.0/OWLPrime and OWLSIF/pD*, which are implemented in Oracle database³. The most advanced OWL profiles are OWL LD that focuses on Linked Data and OWL-RL that is an official standard.

IV. OWL-P

In this section, we present an empirical study of OWL profiles, OWL-P description, comparison to other profiles and support of our proposal in existing RDF graph stores. Before specifying what elements should be supported by OWL-P we analyzed data snapshot 2015 and identified the presence of OWL vocabulary terms.

A. Empirical study about OWL terms

In this Subsection, we analyze representative datasets from the RDF world.

We choose datasets based on Linked Open Data (LOD) Cloud [29]. We gathered the datasets from the Web in three ways: datahub.io dataset catalog⁴, public-lod@w3.org mailing⁵, Billion Triple Challenge 2012⁶. The snapshot is built by LDSpider [30]. The total number of data sets is 1026. Current dataset are classified into the categories: social networking (51%), government (19%), publications (10%), life sciences (7%), user-generated content (6%), cross-domain (4%), media (2%) and geographic (1%).

Table I shows which OWL vocabulary terms are used the most frequently. It is not surprising that the most frequently used are RDF(S) terms. The most popular feature of OWL is `owl:sameAs`. It is worth noting that the least used terms are properties, which were introduced in OWL2.

B. OWL2 Profile

In this Subsection we describe our OWL2 profile. Here we discuss which terms of OWL2 should be supported by OWL-P.

To decide which elements of the OWL vocabulary should be supported by OWL-P, we took into account the results in Table

²<http://franz.com/agraph/allegrograph/>

³<http://www.oracle.com/technetwork/database/index.html>

⁴<http://datahub.io/group/lodcloud>

⁵<https://lists.w3.org/Archives/Public/public-lod/>

⁶<http://km.aifb.kit.edu/projects/btc-2012/>

¹<http://jena.apache.org/documentation/inference>

TABLE I
VOCABULARY TERMS USED IN LOD SNAPSHOT 2015

Vocabulary terms	Occurrence	Vocabulary terms	Occurrence
rdf:type	25695302	owl:differentFrom	784
owl:sameAs	3967150	owl:TransitiveProperty	267
rdfs:subClassOf	1339391	owl:equivalentProperty	201
owl:minCardinality	455203	owl:SymmetricProperty	194
owl:maxCardinality	257371	owl:AllDifferent	111
owl:allValuesFrom	126330	owl:qualifiedCardinality	109
rdfs:domain	111865	owl:InverseFunctionalProperty	94
rdfs:range	59252	owl:propertyChainAxiom	68
owl:unionOf	53735	owl:AllDisjointClasses	21
owl:ObjectProperty	40330	owl:qualifiedMinCardinality	20
owl:equivalentClass	29708	owl:AllDisjointProperties	13
owl:DatatypeProperty	27471	owl:targetValue	11
owl:cardinality	23910	owl:hasKey	5
rdfs:subPropertyOf	13416	owl:propertyDisjointWith	4
owl:someValuesFrom	4446	owl:hasSelf	3
owl:disjointWith	3743	owl:qualifiedMaxCardinality	2
owl:FunctionalProperty	3730	owl:assertionProperty	0
owl:intersectionOf	2681	owl:AsymmetricProperty	0
owl:hasValue	1877	owl:disjointUnionOf	0
owl:inverseOf	1341	owl:IrreflexiveProperty	0
owl:complementOf	873	owl:sourceIndividual	0
owl:oneOf	853	owl:targetIndividual	0

I. Moreover, we considered a time complexity for detecting a required rule application. Because of the complexity we limit elements of body, $n \leq 3$ and we limit elements of head, $m \leq 4$. Therefore OWL-P drops support for restriction and cardinality classes, class relationships and list-based axioms. The most important impact on complexity belongs to blank nodes (mainly present in the list-based axioms). Inferencing with blank nodes often requires an isomorphism check, for which in general, no polynomial algorithms are known in the context of RDF [31].

OWL-P like OWL-RL do not support cardinality restrictions. Restriction classes terms (i.e. `owl:allValuesFrom`, `owl:someValuesFrom`) are too complicated ($m > 4$). Disjunction, keys and property chains terms are unsupported because they are not blank node free and they use lists.

We propose inference rules for OWL-P in N3Logic, because it is a minimal extension to the RDF data model and it can be used for logic and data. Following [32], we define a rule definition below.

Definition 7 (Rule): A rule R has a form $B_1 \wedge \dots \wedge B_n \rightarrow H_1 \wedge \dots \wedge H_m$ where $B_1 \wedge \dots \wedge B_n$ is a *body* of rule and $H_1 \wedge \dots \wedge H_m$ is a *head* of rule.

A body of N3Logic rule and a head of N3Logic rule are written in form of formula (Definition 3). Between the body and the head is the implication (Definition 4).

Taking into consideration terms occurrences and complexity, OWL-P supports the following RDF(S) 1.1 and OWL 2.0 features: RDF(S) terms (`rdf:type`, `rdfs:subClassOf`, `rdfs:subPropertyOf`, `rdfs:domain`, `rdfs:range`), property characteristics (`owl:inverseOf`, `owl:FunctionalProperty`, `owl:InverseFunctionalProperty`, `owl:SymmetricProperty`, `owl:TransitiveProperty`), equivalence between

classes and properties (`owl:equivalentClass`, `owl:equivalentProperty`, `owl:disjointWith`, `owl:propertyDisjointWith`), individual equality and inequality (`owl:sameAs`, `owl:differentFrom`).

We assume that RDF and RDF Schema are the subsets of OWL-P so we support this terms in OWL-P. Properties such `owl:inverseOf`, `owl:sameAs` and `owl:differentFrom` are widely used so we decide to add them to OWL-P. For similar reasons, we support OWL1 property characteristics, such as functional, inverse functional, symmetric, and transitive. Supporting terms that describe the (in)equivalence between classes and properties do not cost much (see rules: `cax-eqc1`, `cax-dw`, `cax-eqc2`, `scm-eqc1`, `scm-eqc2`, `scm-eqp1`, `scm-eqp2`, and `prp-pdw`).

Fig. 1 present inference rules for properties and Fig. 2 present inference rules for classes. In [1], RDF(S) rules that complement OWL-P rules are presented. A syntactic correctness of rules are tested in reasoning engines such as FuXi⁷ and cwm⁸.

C. OWL-P and different approaches

In this Subsection we compare our proposal to other languages based on OWL.

In Table II we analyze existing proposals for different OWL2 profiles. OWL-P is simpler than OWL-RL and OWL LD. Our OWL2 profile drops support for restriction and cardinality classes, class relationships, list-based axioms and some of property characteristics. OWL-P supports more terms than RDFS++ and L2.

The inference rules that we present in this Section are the basis of the deductive RDF graph stores.

⁷<https://github.com/RDFLib/FuXi>

⁸<http://www.w3.org/2000/10/swap/doc/cwm.html>

TABLE II
COMPARISON OF OWL PROFILES

Vocabulary terms	OWL-P	RDFS++ [23]	L2 [24]	RDF 3.0 [25]	pD* [26]	OWL LD [27]	OWL 2 RL [28]
owl:AllDifferent	☒	☒	☒	☒	☒	☒	☒
owl:AllDisjointClasses	☒	☒	☒	☒	☒	☒	☒
owl:AllDisjointProperties	☒	☒	☒	☒	☒	☒	☒
owl:allValuesFrom	☒	☒	☒	☒	☒	☒	☒
owl:assertionProperty	☒	☒	☒	☒	☒	☒	☒
owl:AsymmetricProperty	☒	☒	☒	☒	☒	☒	☒
owl:cardinality	☒	☒	☒	☒	☒	☒	☒*
owl:complementOf	☒	☒	☒	☒	☒	☒	☒
owl:DatatypeProperty	☒	☒	☒	☒	☒	☒	☒
owl:differentFrom	☒	☒	☒	☒	☒	☒	☒
owl:disjointUnionof	☒	☒	☒	☒	☒	☒	☒
owl:disjointWith	☒	☒	☒	☒	☒	☒	☒
owl:equivalentClass	☒	☒	☒	☒	☒	☒	☒
owl:equivalentProperty	☒	☒	☒	☒	☒	☒	☒
owl:FunctionalProperty	☒	☒	☒	☒	☒	☒	☒
owl:hasKey	☒	☒	☒	☒	☒	☒	☒
owl:hasSelf	☒	☒	☒	☒	☒	☒	☒
owl:hasValue	☒	☒	☒	☒	☒	☒	☒
owl:intersectionof	☒	☒	☒	☒	☒	☒	☒
owl:InverseFunctionalProperty	☒	☒	☒	☒	☒	☒	☒
owl:inverseOf	☒	☒	☒	☒	☒	☒	☒
owl:IrreflexiveProperty	☒	☒	☒	☒	☒	☒	☒
owl:maxCardinality	☒	☒	☒	☒	☒	☒	☒*
owl:minCardinality	☒	☒	☒	☒	☒	☒	☒*
owl:ObjectProperty	☒	☒	☒	☒	☒	☒	☒
owl:oneOf	☒	☒	☒	☒	☒	☒	☒
owl:propertyChainAxiom	☒	☒	☒	☒	☒	☒	☒
owl:propertyDisjointWith	☒	☒	☒	☒	☒	☒	☒
owl:qualifiedCardinality	☒	☒	☒	☒	☒	☒	☒*
owl:qualifiedMaxCardinality	☒	☒	☒	☒	☒	☒	☒*
owl:qualifiedMinCardinality	☒	☒	☒	☒	☒	☒	☒*
owl:sameAs	☒	☒	☒	☒	☒	☒	☒
owl:someValuesFrom	☒	☒	☒	☒	☒	☒	☒
owl:sourceIndividual	☒	☒	☒	☒	☒	☒	☒
owl:SymmetricProperty	☒	☒	☒	☒	☒	☒	☒
owl:targetIndividual	☒	☒	☒	☒	☒	☒	☒
owl:targetValue	☒	☒	☒	☒	☒	☒	☒
owl:TransitiveProperty	☒	☒	☒	☒	☒	☒	☒
owl:unionof	☒	☒	☒	☒	☒	☒	☒
rdfs:domain	☒	☒	☒	☒	☒	☒	☒
rdfs:range	☒	☒	☒	☒	☒	☒	☒
rdfs:subClassOf	☒	☒	☒	☒	☒	☒	☒
rdfs:subPropertyOf	☒	☒	☒	☒	☒	☒	☒

* partial supported

Definition 8 (Deductive RDF graph store): A deductive RDF graph store is an entity which store RDF triples and can generate new ones under certain conditions through deduction or inference.

A deductive RDF graph store can answer queries about the combined given and inferred triples. In Table III we present OWL-P support in deductive RDF graph stores. Most OWL-P terms are supported in presented RDF graph stores. The `owl:propertyDisjointWith` has the worst support. Oracle 12c, Pellet and Stardog fully support OWL-P. Not all Jena reasoners support OWL-P.

V. EXPERIMENTS

All experiments have been executed on Intel Xeon Processor E5-2670v2 (2 processors, 20 cores, 40 threads), 128GB of RAM (clock speed: 1866MHz), and a HDD 600GB SAS 10Krpm. We have been used Red Hat 4.4.7-4 (kernel version 2.6.32-431.el6.x86_64).

We gathered the datasets from the Web in two ways:

- 1) crawled data,
- 2) ontologies:
 - a) ChEBI [33],
 - b) Gene Ontology [34],
 - c) MeSH Ontology [35].

The first dataset was generated in LDSpider [30]. The dataset mainly concerns FOAF information because we used FOAF URIs in the seed file. The second group of datasets are ontologies and vocabularies [33], [34], [35]. ChEBI [33] is dictionary of molecular entities focused on chemical compounds. Gene Ontology [34] is controlled vocabulary describe gene and protein roles in cells that is accumulating and changing. MeSH Ontology [35] is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences.

In Table IV we present characteristics of OWL-P inferenc-

TABLE III
OWL-P SUPPORT

Deductive RDF graph stores	df	dw	ec	ep	if	pdw	sa	t	d	r	sco	spo
4rs (4store)	☒	☒	☒	☒	☒	☒	☒	☑	☑	☑	☑	☑
AllegroGraph	☒	☒	☒	☒	☑	☒	☑	☑	☑	☑	☑	☑
Blazegraph	☒	☒	☒	☒	☒	☒	☒	☑	☑	☑	☑	☑
Jena	☑*	☑	☑	☑	☑	☒	☑*	☑	☑	☑	☑	☑
Mulgara	☒	☒	☒	☒	☒	☒	☑	☑	☑	☑	☑	☑
Ontotext OWLIM	☒	☒	☒	☒	☑	☒	☑	☑	☑	☑	☑	☑
Oracle 11g	☑	☑	☑	☑	☒	☒	☑	☑	☑	☑	☑	☑
Oracle 12c	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Pellet	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Sesame	☒	☒	☒	☒	☒	☒	☒	☑	☑	☑	☑	☑
Stardog	☑	☑	☑	☑	☒	☒	☑	☑	☑	☑	☑	☑
Virtuoso	☒	☒	☒	☒	☑	☒	☑	☒	☒	☒	☑	☑

df – owl:differentFrom

dw – owl:disjointWith

ec – owl:equivalentClass

ep – owl:equivalentProperty

if – owl:inverseOf

pdw – owl:propertyDisjointWith

sa – owl:sameAs

t – rdf:type

d – rdfs:domain

r – rdfs:range

sco – rdfs:subClassOf

spo – rdfs:subPropertyOf

* supported by full and mini reasoners

eq-ref	{?S ?P ?O}	=> {?S owl:sameAs ?S. ?P owl:sameAs ?P. ?O owl:sameAs ?O}.	cax-eqcl	{?A owl:equivalentClass ?B. ?X rdf:type ?A}	=> {?X rdf:type ?B}.
eq-sym	{?S owl:sameAs ?O}	=> {?O owl:sameAs ?S}.	cax-eqc2	{?A owl:equivalentClass ?B. ?X rdf:type ?B}	=> {?X rdf:type ?A}.
eq-trans	{?Q owl:sameAs ?R. ?R owl:sameAs ?P}	=> {?Q owl:sameAs ?P}.	cax-dw	{?A owl:disjointWith ?B. ?X rdf:type ?A. ?X rdf:type ?B}	=> {@false}
eq-rep-s	{?S owl:sameAs ?S2. ?S ?P ?O}	=> {?S2 ?P ?O}.	scm-cls	{?C rdf:type owl:Class}	=> {?C rdfs:subClassOf ?C. ?C rdfs:subClassOf owl:Thing. ?C owl:equivalentClass ?C. owl:Nothing rdfs:subClassOf ?C}.
eq-rep-p	{?P owl:sameAs ?P2. ?S ?P ?O}	=> {?S ?P2 ?O}.	scm-eqcl	{?A owl:equivalentClass ?B}	=> {?A rdfs:subClassOf ?B. ?B rdfs:subClassOf ?A}.
eq-rep-o	{?O owl:sameAs ?O2. ?S ?P ?O}	=> {?S ?P ?O2}.	scm-eqc2	{?A rdfs:subClassOf ?B. ?B rdfs:subClassOf ?A}	=> {?A owl:equivalentClass ?B}. {?P rdfs:subPropertyOf ?P. owl:ObjectProperty}
eq-diff1	{?Q owl:sameAs ?R. ?Q owl:differentFrom ?R}	=> {@false}.	scm-op	{?P rdf:type owl:ObjectProperty}	=> {?P owl:equivalentProperty ?P}. {?P rdfs:subPropertyOf ?P. owl:DatatypeProperty}
prp-eqp1	{?P1 owl:equivalentProperty ?P2. ?Q ?P1 ?R}	=> {?Q ?P2 ?R}.	scm-dp	{?P rdf:type owl:DatatypeProperty}	=> ?P owl:equivalentProperty ?P}.
prp-eqp2	{?P1 owl:equivalentProperty ?P2. ?Q ?P2 ?R}	=> {?Q ?P1 ?R}.	scm-eqp1	{?P owl:equivalentProperty ?R}	=> {?P rdfs:subPropertyOf ?R. ?R rdfs:subPropertyOf ?P}.
prp-pdw	{?P1 owl:propertyDisjointWith ?P2. ?Q ?P1 ?R. ?Q ?P2 ?R}	=> {@false}.	scm-eqp2	{?P rdfs:subPropertyOf ?R. ?R rdfs:subPropertyOf ?P}	=> {?P owl:equivalentProperty ?R}.
prp-inv1	{?P1 owl:inverseOf ?P2. ?Q ?P1 ?R}	=> {?R ?P2 ?Q}.			
prp-inv2	{?P1 owl:inverseOf ?P2. ?Q ?P2 ?R}	=> {?R ?P1 ?Q}.			
prp-fp	{?P rdf:type owl:FunctionalProperty. ?X ?P ?Y1. ?X ?P ?Y2}	=> {?Y1 owl:sameAs ?Y2}.			
prp-ifp	{?P rdf:type owl:InverseFunctionalProperty. ?X1 ?P ?Y. ?X2 ?P ?Y. }	=> {?X1 owl:sameAs ?X2}.			
prp-symp	{?P rdf:type owl:SymmetricProperty. ?X ?P ?Y. }	=> {?Y ?P ?X}.			
prp-trp	{?P rdf:type owl:TransitiveProperty. ?X ?P ?Y. ?Y ?P ?Z. }	=> {?X ?P ?Z}.			

Fig. 1. OWL-P inference rules for properties

Fig. 2. OWL-P inference rules for classes

ing. To execute our rules we used EYE [36]. The table shows terms that are the most common and available in OWL-P (cf. Subsection IV-A). The results show that the largest increase belongs to ChEBI and the slightest increase belongs to the crawled data. This result is expected, because in the crawled data, the occurrence of OWL-P terms are the smallest.

VI. CONCLUSIONS AND FUTURE WORK

This paper defines how knowledge and logic might be handled on the Semantic Web environment. We present an OWL-P that is a lightweight profile of OWL2. We propose inference rules for our approach. All rules are tested in reasoning engines. This paper provides a specification of OWL-P which can be more simply and efficiently implemented. Our formalization is based on Notation 3 Logic, which extended

RDF by logical symbols and created a new Semantic Web logic. We analyze existing deductive RDF graph stores in the context of our proposal and show that in most software they support OWL-P without any changes.

Future work will focus on preparing OWL-P rules expressed in popular inference rule syntaxes, such as RuleML, and RIF. Moreover, we would like to examine the relationship between our solution and SPIN, the language that allows to create constraints on Semantic Web models. Another challenge is to check all the possible sources of inconsistency in an ontology.

ACKNOWLEDGMENTS

This publication has received financial support from the Polish Ministry of Science and Higher Education under subsidy for maintaining the research potential of the Faculty of Mathematics and Informatics, University of Bialystok.

Moreover, this research was supported by the Computer Center of University of Bialystok grant (GO-027).

TABLE IV
INFERRING CHARACTERISTICS

Characteristics	crawler	ChEBI	GO	MeSH
file size (before) [B]	463972372	15452129	88378906	34494110
file size (after) [B]	1743903440	79716934	443224701	145644585
ratio	3.76	5.16	5.02	4.22
inferencing times [s]	381049.45	14824.62	525513.59	40092.12
triples (before)	630799	259913	1571117	566933
triples (after)	1097421	399747	2380789	646583
ratio	1.74	1.54	1.52	1.14
rdf:type (before)	307139	36682	248474	17076
rdf:type (after)	307139	110073	745489	96726
ratio	1.00	3.00	3.00	5.66
owl:sameAs (before)	15	0	0	0
owl:sameAs (after)	15	0	0	0
ratio	1.00	n/a	n/a	n/a
rdfs:subClassOf (before)	0	50197	95778	2129
rdfs:subClassOf (after)	0	50199	95778	2211
ratio	n/a	1.00	1.00	1.04
owl:equivalentClass (before)	0	0	11899	0
owl:equivalentClass (after)	0	0	11901	2
ratio	n/a	n/a	1.00	n/a
rdfs:subPropertyOf (before)	0	0	25	0
rdfs:subPropertyOf (after)	0	0	27	0
ratio	n/a	n/a	1.08	n/a

REFERENCES

- [1] D. Tomaszuk, "Inference rules for RDF(S) and OWL in N3Logic," *arXiv preprint arXiv:1601.02650*, 2016.
- [2] V. Kolovski, Z. Wu, and G. Eadon, "Optimizing enterprise-scale OWL 2 RL reasoning in a relational database system," *The Semantic Web - ISWC 2010*, pp. 436–452, 2010. doi: 10.1007/978-3-642-17746-0_28. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-17746-0_28
- [3] D. Wood, M. Lanthaler, and R. Cyganiak, "RDF 1.1 Concepts and Abstract Syntax," World Wide Web Consortium, W3C Recommendation, Feb. 2014. [Online]. Available: <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>
- [4] D. Tomaszuk, L. Skonieczny, and D. Wood, "RDF Graph Partitions: A Brief Survey," in *BDAS*, ser. Communications in Computer and Information Science, vol. 521. Springer, 2015. doi: 10.1007/978-3-319-18422-7_23. ISBN 978-3-319-18421-0 pp. 256–264. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-18422-7_23
- [5] D. Arndt, R. Verborgh, J. De Roo, H. Sun, E. Mannens, and R. Van de Walle, "Semantics of Notation3 Logic: A solution for implicit quantification," in *Proceedings of the 9th International Web Rule Symposium*, Aug. 2015. doi: 10.1007/978-3-319-21542-6_9. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-21542-6_9
- [6] B. Harold, M. Dean, B. Grosf, M. Sintek, B. Spencer, S. Tabet, and G. Wagner, "FOL RuleML: The First-Order Logic Web Language," Tech. Rep., Nov. 2004. [Online]. Available: <http://ruleml.org/fol/>
- [7] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosf, M. Dean *et al.*, "SWRL: A semantic web rule language combining OWL and RuleML," World Wide Web Consortium, W3C Member Submission, May 2004. [Online]. Available: <http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/>
- [8] M. Kifer, "Rule interchange format: The framework," in *Web reasoning and rule systems*. Springer, 2008, pp. 1–11. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-88737-9_1
- [9] M. Kifer and H. Boley, "RIF Overview," World Wide Web Consortium, W3C Working Draft, Oct. 2009. [Online]. Available: <https://www.w3.org/TR/2009/WD-rif-overview-20091001/>
- [10] N. Bassiliades and I. Vlahavas, "R-device: A deductive RDF rule language," in *Rules and Rule Markup Languages for the Semantic Web*. Springer, 2004, pp. 65–80. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-30504-0_6
- [11] M. Sintek and S. Decker, "TRIPLE-An RDF Query, Inference, and Transformation Language," in *INAP*, 2001. doi: 10.1007/3-540-48005-6_28 pp. 47–56. [Online]. Available: http://dx.doi.org/10.1007/3-540-48005-6_28
- [12] H. Knublauch, J. A. Hendle, and K. Idehen, "SPIN - Overview and Motivation," World Wide Web Consortium, W3C Member Submission, Feb. 2011. [Online]. Available: <http://www.w3.org/Submission/2011/SUBM-spin-overview-20110222/>
- [13] B. Harold, T. Athan, A. Paschke, A. Giurca, N. Bassiliades, G. Governatori, M. Palmirani, A. Wyner, G. Zou, and Z. Zhao, "Specification of Deliberation RuleML 1.01," Tech. Rep., 2012. [Online]. Available: http://wiki.ruleml.org/index.php/Specification_of_Deliberation_RuleML_1.01
- [14] B. Parsia, S. Rudolph, M. Krötzsch, P. Patel-Schneider, and P. Hitzler, "OWL 2 Web Ontology Language Primer (Second Edition)," World Wide Web Consortium, W3C Recommendation, Dec. 2012. [Online]. Available: <http://www.w3.org/TR/2012/REC-owl2-primer-20121211/>
- [15] D. Reynolds, M. Kifer, A. Polleres, H. Boley, A. Paschke, and G. Hallmark, "RIF Core Dialect (Second Edition)," World Wide Web Consortium, W3C Recommendation, Feb. 2013. [Online]. Available: <http://www.w3.org/TR/2013/REC-rif-core-20130205/>
- [16] M. Kifer and H. Boley, "RIF Basic Logic Dialect (Second Edition)," World Wide Web Consortium, W3C Recommendation, Feb. 2013. [Online]. Available: <http://www.w3.org/TR/2013/REC-rif-bld-20130205/>
- [17] C. d. S. Marie, A. Paschke, and G. Hallmark, "RIF Production Rule Dialect (Second Edition)," World Wide Web Consortium, W3C Recommendation, Feb. 2013. [Online]. Available: <http://www.w3.org/TR/2013/REC-rif-prd-20130205/>
- [18] A. Polleres, M. Kifer, and H. Boley, "RIF Datatypes and Built-Ins 1.0 (Second Edition)," World Wide Web Consortium, W3C Recommendation, Feb. 2013. [Online]. Available: <http://www.w3.org/TR/2013/REC-rif-dtb-20130205/>
- [19] A. Horn, "On sentences which are true of direct unions of algebras," *The Journal of Symbolic Logic*, vol. 16, no. 01, pp. 14–21, 1951. doi: 10.2307/2268661. [Online]. Available: <http://dx.doi.org/10.2307/2268661>
- [20] M. Kifer and G. Lausen, "F-logic: a higher-order language for reasoning about objects, inheritance, and scheme," in *ACM SIGMOD Record*, vol. 18, no. 2. ACM, 1989. doi: 10.1145/66926.66939 pp. 134–146. [Online]. Available: <http://dx.doi.org/10.1145/66926.66939>
- [21] B. McBride, "Jena: A semantic web toolkit," *IEEE Internet computing*, no. 6, pp. 55–59, 2002. doi: 10.1109/MIC.2002.1067737. [Online]. Available: <http://dx.doi.org/10.1109/MIC.2002.1067737>
- [22] T. Rattanasawad, K. R. Saikaew, M. Buranarach, and T. Supnithi, "A review and comparison of rule languages and rule-based inference engines for the Semantic Web," in *Computer Science and Engineering Conference (ICSEC), 2013 International*. IEEE,

2013. doi: 10.1109/ICSEC.2013.6694743 pp. 1–6. [Online]. Available: <http://dx.doi.org/10.1109/ICSEC.2013.6694743>
- [23] “AllegroGraph 6.2.2 Reasoner Tutorial,” accessed: 2017-07-06. [Online]. Available: <https://franz.com/agraph/support/documentation/current/reasoner-tutorial.html>
- [24] F. Fischer, G. Unel, B. Bishop, and D. Fensel, “Towards a scalable, pragmatic knowledge representation language for the Web,” in *Perspectives of Systems Informatics*. Springer, 2010, pp. 124–134. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-11486-1_11
- [25] J. Hendler, “RDFS 3.0,” in *W3C Workshop – RDF Next Steps*. World Wide Web, 2010. [Online]. Available: <https://www.w3.org/2009/12/rdf-ws/papers/ws31>
- [26] H. J. ter Horst, “Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 3, no. 2, pp. 79–115, 2005. doi: 10.1016/j.websem.2005.06.001. [Online]. Available: <http://dx.doi.org/10.1016/j.websem.2005.06.001>
- [27] B. Glimm, A. Hogan, M. Krotzsch, and A. Polleres, “OWL LD: Entailment Ruleset and Implementational Notes.” [Online]. Available: <http://semanticweb.org/OWLLD/>
- [28] B. Motik, B. C. Grau, I. Horrocks, A. Fokoue, and Z. Wu, “OWL 2 Web Ontology Language Profiles (Second Edition),” World Wide Web Consortium, W3C Recommendation, Dec. 2012. [Online]. Available: <http://www.w3.org/TR/2012/REC-owl2-profiles-20121211/>
- [29] M. Schmachtenberg, C. Bizer, and H. Paulheim, “Adoption of the linked data best practices in different topical domains,” in *The Semantic Web–ISWC 2014*. Springer, 2014, pp. 245–260. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11964-9_16
- [30] R. Isele, J. Umbrich, C. Bizer, and A. Harth, “LDSpider: An open-source crawling framework for the Web of Linked Data,” in *Proceedings of 9th International Semantic Web Conference (ISWC 2010) Posters and Demos*, 2010.
- [31] A. Hogan, M. Arenas, A. Mallea, and A. Polleres, “Everything You Always Wanted to Know About Blank Nodes,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 27, no. 1, 2014. doi: 10.1016/j.websem.2014.06.004. [Online]. Available: <http://dx.doi.org/10.1016/j.websem.2014.06.004>
- [32] E. Liarou, S. Idreos, and M. Koubarakis, “Evaluating conjunctive triple pattern queries over large structured overlay networks,” in *The Semantic Web–ISWC 2006*. Springer, 2006, pp. 399–413. [Online]. Available: http://dx.doi.org/10.1007/11926078_29
- [33] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, “ChEBI: a database and ontology for chemical entities of biological interest,” *Nucleic acids research*, vol. 36, no. suppl 1, pp. D344–D350, 2008. doi: 10.1093/nar/gkm791. [Online]. Available: <http://dx.doi.org/10.1007/10.1093/nar/gkm791>
- [34] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, “Gene Ontology: tool for the unification of biology,” *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000. doi: 10.1038/75556. [Online]. Available: <http://dx.doi.org/10.1038/75556>
- [35] C. E. Lipscomb, “Medical subject headings (MeSH),” *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.
- [36] R. Verborgh and J. De Roo, “Drawing conclusions from linked data on the web: The EYE reasoner,” *IEEE Software*, vol. 32, no. 3, pp. 23–27, 2015. doi: 10.1109/MS.2015.63. [Online]. Available: <http://dx.doi.org/10.1109/MS.2015.63>

8th International Workshop on Advances in Semantic Information Retrieval

THE International Workshop on Advances in Semantic Information Retrieval is organized as an event within FedCSIS. In 2018, we are running our eighth workshop. It is gaining popularity among researchers from Europe and Asia.

We are doing our best to follow the changes in the area of semantic information retrieval and making the necessary adjustments in the set of topics of interests. We shift the focus of this workshop on the most challenging problems. The ASIR'18 workshop will continue to maintain high standards of quality and organization, set in the previous years.

Characterizing the current tendency in development of semantic technologies, we point out that recent advances form a solid basis for a variety of methods and instruments used in multimedia information retrieval, knowledge representation, discovery and analysis. They influence the way and form of representing documents in the memory of computers, approaches to analyze documents, and techniques to mine and retrieve knowledge. In its turn, gathered knowledge is used to build problem domain models and support decision making. The abundance of video, voice and speech data also raises new challenging problems in developing multimedia information retrieval systems.

We believe that the ASIR'18 workshop will facilitate discussions of new research results in this area, and will serve as a meeting place for researchers from all over the world. Our aim is to create an atmosphere of friendship and cooperation for everyone, interested in computational linguistics, data-driven decision making, data analytics and semantic information retrieval. We welcome all interested researchers to join this event.

TOPICS

The topics and areas include but not limited to:

- Data analytics.
- Data-driven decision making.
- Domain-specific semantic applications.
- Evaluation methodologies for semantic search and retrieval.
- Knowledge representation and management.
- Models for document representation.
- Natural language semantic processing.
- Ontology for semantic information retrieval.
- Ontology alignment, mapping and merging.
- Query interfaces.
- Searching and ranking.
- Semantic multimedia retrieval.
- Visualization of retrieved results.

EVENT CHAIRS

- **Klyuev, Vitaly**, University of Aizu, Japan
- **Mozgovoy, Maxim**, University of Aizu, Japan

PROGRAM COMMITTEE

- **Dobrynin, Vladimir**, Saint Petersburg State University, Russia
- **Goczyła, Krzysztof**, Gdansk University of Technology, Poland
- **Gutnova, Alina**, North Ossetian State University, Russia
- **Haralambous, Yannis**, Institut Telecom - Telecom Bretagne, France
- **Homenda, Wladyslaw**, Warsaw University of Technology, Poland
- **Janusz, Andrzej**, University of Warsaw, Poland
- **Jin, Qun**, Waseda University, Japan
- **Kotets, Alexey**, North Ossetian State University, Russia
- **Lai, Cristian**, CRS4, Italy
- **Makarenko, Maria**, North Ossetian State University, Russia
- **Minasyan, David**, North Ossetian State University, Russia
- **Nalepa, Grzegorz J.**, AGH University of Science and Technology, Poland
- **Pyshkin, Evgeny**, University of Aizu, Japan
- **Shtykh, Roman**, CyberAgent Inc., Japan
- **Śluzek, Andrzej**, Khalifa University, United Arab Emirates
- **Suárez-Figueroa, Mari Carmen**, Ontology Engineering Group, School of Computer Science at Universidad Politécnica de Madrid, Spain
- **Tadeusiewicz, Ryszard**, AGH University of Science and Technology, Poland
- **Vacura, Miroslav**, University of Economics, Czech Republic
- **Zadrozny, Sławomir**, Systems Research Institute of Polish Academy of Sciences, Poland
- **Ławrynowicz, Agnieszka**, Poznan University of Technology, Poland

A Chatbot Based On AIML Rules Extracted From Twitter Dialogues

Hiroshi Yamaguchi, Maxim Mozgovoy
The University of Aizu
Tsuruga, Ikki-machi, Aizuwakamatsu, Fukushima,
965-8580 Japan
{m5201115, mozgovoy}@u-aizu.ac.jp

Anna Danielewicz-Betz
Ludwig-Maximilian University
EU Business School
Munich, Germany
anna.danielewicz_betz@euruni.edu

Abstract—A chat dialogue system, a chatbot, or a conversational agent is a computer program designed to hold a conversation using natural language. Many popular chat dialogue systems are based on handcrafted rules, written in Artificial Intelligence Markup Language (AIML). However, a manual design of rules requires significant efforts, as in practice most chatbots require hundreds if not thousands of rules. This paper presents the method of automated extraction of AIML rules from real Twitter conversation data. Our preliminary experimental results show the possibility of obtaining natural-language conversation between the user and a dialogue system without the necessity of handcrafting its knowledgebase.

INTRODUCTION

A *CHAT* dialogue system or a *conversational agent* is a computer program designed to hold conversations in a human-like way and ideally “understand” the user’s intent [1]. In general, chat dialogue systems can be categorised into two types: *task-oriented* systems that are used to assist the user in completing various tasks within a specified domain, and *open-domain* systems that aim at performing a natural conversation with the user [2]. Task-oriented chatbots are deployed in various business settings, such as social media marketing (personalised shopping, simplified buying procedures, customer help, statbots), typically to assist if not replace human customer service in live chats.

Conversational chatbots, on the other hand, may serve, e.g., as alternative interlocutors in healthcare or simply to keep lonely people company. Furthermore, open-domain conversational agents are good testbeds for the development and evaluation of “social” interfaces, deployable in a wide range of applications [3].

Many dialogue systems are rule-based, and one of the most popular mechanisms of representing rules is AIML (Artificial Intelligence Markup Language). AIML is a simple XML-based markup language that gained popularity after being used in the successful dialogue system A.L.I.C.E. [4] that won the Loebner Prize three times. The main drawback of AIML-based systems lies in the large number of rules required to imitate a natural conversation, especially in the case of open-domain systems. Therefore, an AIML-based dialogue system requires considerable

manual effort to describe its knowledgebase, leading to an expensive and error-prone development process.

The goals of the present research are to:

- develop a method for AIML rules generation on the basis of existing conversational corpus;
- address approximate matching and context analysis;
- test the chosen approach using Japanese Twitter as a corpus of natural dialogues;
- evaluate the performance of the resulting chatbot.

Perhaps, the most attractive feature of rule-based systems is their simplicity. However, the need to design numerous rules can become a major obstacle in practice. We intend to show that it is possible to reduce the amount of human effort by automating the rule generation process, using a large dataset of authentic human conversations, such as Twitter dialogues. Certain steps of this process were demonstrated in our earlier system [5], but it notably lacked the ability to track the context of conversation and approximate matching capabilities.

TWITTER AS A CORPUS OF CONVERSATIONS

The Internet serves as a vast corpus of conversational data. One can assume that Twitter dialogues come relatively close to informal daily conversations. There is, moreover, an API available to retrieve individual tweets and tweet streams [6]. This observation motivated us to use Twitter as a source of dialogues that can be converted into AIML rules.

For the previous version of the system [5], we retrieved a dataset of tweets posted between October 2016 and April 2017, using Streaming API [7]. Individual tweets were tagged with several attributes, including unique tweet ID, tweet language, timestamp, and in-reply-to field. For the present work, we used Rest API to extract replies to the tweets already present in our collection. Our goal was to obtain chains of three consecutive dialogue lines: the original tweet, a reply to the tweet, and a reply to the reply. As a result, we gathered a corpus of 49,971 dialogues of three lines or longer, and extracted 614,271 triples from the corpus.

CONTEXTUALISED AIML RULES

AIML is based on XML, and thus consists of hierarchically organised elements. Individual “units of knowledge” are known as *categories* in AIML. Each category should define at least two compulsory elements: a *pattern* that contains a sample input, and a *template* that contains the corresponding response of the chatbot [8].

In the following example, if the user inputs おはよう! (“Morning!”), the bot should reply おはようございます (“Good morning”):

```
<category>
  <pattern>おはよう!</pattern>
  <template>おはようございます。</template>
</category>
```

AIML syntax supports approximate matching with wildcard symbols and a mechanism of redirection, used to handle different situations with the same rules. AIML also allows specifying the context where the given rule is applicable, and thus keeping dialogues coherent. We rely on this capability when converting Twitter dialogues into AIML rules. The resulting system uses the rules including all three elements:

```
<category>
  <context>おはよう!</context>
  <pattern>おはようございます。</pattern>
  <template>今日はいい天気ですね。</template>
</category>
```

Here the bot will reply 今日はいい天気ですね。 (“It’s a good weather today”) only if the two preceding dialogue lines were おはよう! (“Morning!”) and おはようございます (“Good morning”).

The set of AIML rules forming a chatbot’s knowledgebase is processed with an *AIML interpreter*, responsible for actual dialogues with the user. We use an in-house developed interpreter that implements approximate matching and tokenization of Japanese texts.

CONVERTING DIALOGUES INTO AIML RULES

The process of converting the raw tweet dataset into a set of AIML rules consists of the following steps.

Preprocessing: Raw tweet data contain numerous messages without any conversational meaning that have to be considered irrelevant for our purposes. These tweets typically consist of hyperlinks, hashtags and/or user names, or contain no Japanese characters. We remove such tweets from the source collection.

Normalisation: Each element in our collection contains three consecutive dialogue lines that are to be mapped to the AIML tags <context>, <pattern> and <template>.

Our system attempts to identify the best matching context and pattern for the current situation using TF-IDF approach [9]. This process requires tokenization into individual morphemes, and stop-words removal. This is done with the help of Japanese morphological analyser MeCab [10] which splits the text into individual part-of-speech tagged morphemes. We use part-of-speech tags to eliminate non-significant morphemes, such as auxiliary verbs, postpositional particles, conjunctions, and pre-noun adjectivals.

Rule Generation. Each triple is transformed into an individual AIML rule. Triple elements are mapped to the AIML tags <context>, <pattern> and <template>.

INTERPRETING AIML

The semantics of individual AIML elements is well documented, which makes it possible to develop a universal AIML interpreter, able to serve as a chatbot powered with any given set of AIML rules. Indeed, some interpreters, such as Program AB [11] or pyAIML [12] are freely available.

Since we wanted to make use of TF-IDF based approximate matching, we had to implement our own AIML interpreter. The present AIML specification supports approximate matching, but this capability is based on wildcard characters rather than a text similarity analysis. On the other hand, we had to support the most basic AIML syntax that relies on <context>, <pattern>, and <template> tags only, so the resulting interpretation algorithm is relatively straightforward.

The current version of the system operates as follows. The chatbot starts a dialogue with a line こんにちは (“Hello”). The user’s reply provides a context/pattern pair that is used to retrieve the next dialogue line of the chatbot (the highest scoring match according to TF-IDF is chosen). The last two replies become the new context and pattern, and the whole process is repeated.

EVALUATION CRITERIA

Different evaluation criteria can be applied, depending on the goals a given chatbot has been created to fulfil or tasks to perform. In other words, the evaluation criteria depend on the metrics applied at conceptual, operational, and qualitative levels. As for quality, here cohesion, cooperation, likeability, engagement, trust, reduction of frustration or ability to comment and provide feedback play a role. Cognitive linguistic quality criteria include, broadly speaking, conversational flow, understanding, and accuracy. Liu et al. [13] refer to task (completion)-focused responses and user satisfaction scores based on “model responses” and “appropriateness” of the proposed response to the conversation at hand, whereby a semantic match — co-occurrence in a given context — has to take place, especially for “informative words”, as opposed to the “common” ones. They introduce a metric that correlates more strongly with human judgement, with the goal to

automatically evaluate how “appropriate” the proposed response is to the conversation, resorting to two approaches: word-based similarity metrics and word-embedding-based similarity metrics. Chakrabarti and Luger [14] refer to a goal-fulfilment map fostering evaluation that is to perform adequately not only in an isolated question-answer exchange, but also in a longer, sustained conversation, with the dialogue agent’s enhanced ability to adhere to context in conversations, to hold a longer conversation, and more closely emulate a human-like conversation. This entails knowing what to say (content), knowing how to express it through a conversation (semantics), and having a standard benchmark to assess conversations (pragmatics-based evaluation, including input that is relevant to the context and within a given domain).

According to [15], adaptation to new information/request (that is, e.g., matching a given speech act while observing conversational rules) is an important quality factor, along with usability connected to effectiveness, efficiency, and satisfaction with contextually-bound goal fulfilment. On the other hand, a conversational agent should not aim at acting human. The evaluation categories encompass performance, manifested in such quality attributes as “robustness to unexpected input”, “appropriateness and ability to perform damage control”, “effective function allocation — provision of escalation channels”; functionality — manifested in “accurate speech synthesis”, “accurate interpretation of commands”, “appropriate degree of formality and accuracy of outputs”, and “execution of requested tasks”. Moreover, in terms of humanity criterion, the interaction should be “convincing, satisfying and natural”, with the chatbot’s ability to “respond to specific questions and to maintain themed discussion”.

Overall, the quality attributes proposed include, among other things, robustness to unexpected input, provision of appropriate escalation channels, ability to maintain “themed discussion”, i.e. within a given domain; being entertaining and engaging, ability to detect meaning and intent, and ability to respond to social cues.

Saygin and Cicekli [16] point out that the principles guiding human-computer conversation may be slightly different from those guiding inter-human communication. They propose that Grice’s “cooperative principle” [17], consisting of conversational maxims, be taken into account when evaluating human-machine communication, albeit in a modified form. In particular, relevance maxim should not be violated by a dialogue system since, contrary to the human intention to change the subject, joke or use a metaphor, this is interpreted as inability to understand input utterances. By contrast, violations of manner have a positive effect on imitating human-like behaviour in that overreactive displays of emotions and impoliteness are normally associated with humans. So violations of relevance tend to create a machine-like effect and those of manner tend to create a human-like effect. Furthermore, violation of quantity maxim creates a machine-like effect since there is a strong

correlation between this maxim and “artificial language use”. As for quality maxim, no strong conclusions were reached since its violations tended to occur together with those of quantity, manner, and especially relevance. Since the deployment of a conversational agent cannot involve any cooperation *per se*, but rather imitation or simulation of thereof, the authors propose that the conversational principle be modified to accommodate human-computer interaction. In human-human conversations, the maxims are regularly flouted on purpose or violated unintentionally, yet this will not necessarily result in a communication breakdown, unlike in human-machine interaction.

Since the business goals of intelligent agents differ from purely conversational purposes (e.g. increased customer satisfaction, personalised solutions), so do the evaluation criteria as users expect intuitive, fast and “valuable” conversations. Such chatbot applications fall, however, outside the scope of the present paper.

EVALUATION AND FINDINGS

We conducted a pilot evaluation test of our chatbot, involving 10 respondents (5 female and 5 male), all speakers of Japanese (6 undergraduate students and 4 older adults aged 29-58). Each person made 3 attempts at chatting with the bot, resulting in total 30 chats (23 of which were 10 lines long on average; and 7 ranging from 17 to 41 lines). The evaluation questions on a 3-point Likert scale were adapted from [16] and answered by each respondent, following the three chat attempts. Due to the convenience sample of 10 respondents with rather limited exposure to the chatbot, as well as varying conversational skills and the negativity bias toward a machine interlocutor, the evaluation findings should be treated as preliminary. For the sake of reading convenience and comprehension, we only provide conversations translated into English in the subsequent sections.

A. Pragmatic Analysis

In pragmatics, the term “adjacency pairs”, in connection to speech acts [18, 19], refers to those turns in conversations that have specific follow-ups, e.g. greeting-greeting, question-answer, invitation-acceptance or apology-acknowledgement. Opening sequences serve to initiate a conversation by means of greetings and small talk (general questions or comments about the weather, sports, etc.); whereas closing sequences signal an ending of a conversation (e.g. okay, all right then, well), followed by repetitions of farewells (okay, goodbye then; okay bye). Openings and closings are more conventionalised than are other parts of the conversation. The term “repair” refers to the clarification of previous intentions or the need of editing a preceding statement, i.e. “fixing” the utterance in some way. Politeness refers to conventionalised ways of conversing in an appropriate way that may involve titles and address forms, being indirect, that is generally avoiding any face threatening acts (see Politeness Principle [20]).

Based on the pragmatic analysis of the 30 chats in our sample dataset, certain tendencies regarding “conversational behaviour” of both the chatbot (B) and that of the users (U) were observed. It can be noted that the chatbot tends to successfully complete adherent speech acts in opening sequences, such as greetings (“Hello” — “Hello”); small talk questions about general well-being (U: “How are you?” — B: “I’m well”, reinforced at times by emoticons, or questions or remarks about the weather (“The weather is hot”).

Generally, it tackles question-answer sequences in a satisfactory way by providing a generic answer (U: “Will you go shopping?”/“Where will you go?” — B: “I don’t know”); by answering a yes/no question (U: “Are you hungry?” — B: “Yes, I am”; U: “Do you like music?” — B: “Yes, I like it [*note emoji*]”) or by giving a more detail answer (U: “Where do you work?” — B: “I work in a factory [*thumbs up emoji*]”). Depending on the domain and question complexity, an attempt at a more elaborate answer may stretch over a number of lines, if not “interrupted” by the user’s impatience and an abrupt change of subject (U: “Do you enjoy painting?” — B: (...) “I’m not good at sketching in five minutes [*emoticons*]. But I’ll get experience [*emoticons*]. I want to improve my skills. I will try”). In most cases, however, the users, not aiming at exploration of a given topic, fail to ask further questions, (e.g., “What kind of music do you like? Do you like jazz?”), and change the topic abruptly.

In addition, the conversational agent simulates emotive reactions (with reference to senses) on a number of occasions, reinforced by punctuation marks, emoticons and/or emoji (“Lean on me” — “I want to pat you on the head” — “Oh... you’re fluffy” — all in one conversational sequence). Unfortunately, the users tend not to follow up on such “emotional vibes”, resorting to dispreferred options (causing a mismatch in speech acts) in their responses, as illustrated by this example:

| B: I love you
| U: No, thank you.

Here the user exhibits outright violation of maxim of relevance, as well as that of politeness.

The chatbot makes rather successful attempts at simulating emotions — frequently reinforced by emoticons/emoji (“Oh, I’m embarrassing”, “Envy”), It also “expresses” concern, but, again, such attempts are not pursued further by the users, reluctant to explore the topic of emotions:

| B: My heart is frozen by your reply
| U: My muscles are aching!

Emotional content is thus commonly met with inappropriate answers, repetition or change of subject.

B. User-Induced Failure in Communication

The following example represents a dispreferred pair, hence the subsequent problems in matching an appropriate response by the bot:

| B: Hello

| U: What’s your name?

A failure to reciprocate a greeting results in communication difficulties as the bot repeats its attempts at greeting (B: “Hello” — U: “What kind of people do you like?” — B: “Hello”). The same applies to conversations initiated by the user: Since the greeting speech act is not performed on numerous occasions, this causes a mismatch in adjacency pairs (a dispreferred second part).

Moreover, the user’s mono-word utterances, such as “OK”, “Who?”, “What?”, “Yes”, “Really” — violating the conversational maxim of quantity [17] — pose a great challenge to the bot searching for appropriate follow-up candidate responses to match its response with, as such common words devoid of context cannot be assigned to any specific thematic domain or speech act adjacency pair (U: “Are you OK?” — B: “Thanks” — U: “What?” — B: “Kittyoppa-”). They tend to lead to a breakdown in communication, especially if used in a context requiring a more elaborate reaction (B: “You’re a liar” — U: “What?”). By simply confirming with “Yes”, for instance, and not exploiting the topic further, the user signals that she/he wishes to finish the interaction (B: “Let’s hang out!” — U: “What are we doing?” — B: “Today?” — U: “Yes”).

As mentioned before, the users also tend to change the subject abruptly (or ask questions that do not make much sense), thus violating the maxim of relevance. This poses immense difficulties for the bot. Simultaneously, the users appear to exhibit negative bias toward the computer system [16], demonstrated in questioning the chatbot’s overall ability to communicate (“You cannot communicate”; “You really don’t understand”; “What are you talking about?”). Consequently, instead of focusing on a conversation at hand, the users switch to the “evaluation mode”, which hinders conversational flow and “confuses” the chatbot entirely. Generally, in the face of any slight difficulty, rather than rephrasing or clarifying their utterance (repairs are not common) or resorting to another conversational domain (since only limited domains of conversation such as greeting, weather, and food are explored) where a conversation might be more successful, the users accuse or reproach the bot (“I don’t understand”/“What?”), thus violating the principles of cooperation and politeness themselves.

The challenging speech act of reprimand-apology is, however, tackled rather well (U: “You’re a fool” — B: “It is regrettable” — U: “I’m sorry”), whereby the ensuing bot’s utterance attempts to pick on the unjust reprimand (“It’s so dangerous that you are strange”), but this rather sophisticated attempt at a further exchange on the topic at hand is not recognised as a follow-up sequence by the user.

In longer conversations (20 lines onward), there is no flow or cohesion that is normally demonstrated in human-human turn taking with smooth transitions from one topic to the other, but rather incohesive utterance chunks divide the whole “conversation” into unrelated parts. This is, again, at least partly due to the users’ inability to lead and steer a

conversation skillfully to a variety of topics, by providing longer utterances, if not some contextual information; or by not recognising the attempts to continue a topic broached upon a couple of lines earlier. Such behaviour does not facilitate further conversation.

Overall, the users appear to violate the quantity, relevance, quality and manner maxims. They answer very briefly, even in single words and discuss a very limited range of topics — mostly food in a repetitive manner, then weather or music, and rarely work (“I want meat”, “I’m thinking what to eat”). Moreover, as mentioned before, the users frequently change the subject in an unexpected way, and then accuse the chatbot of being unable to communicate (“We don’t communicate well”):

B: I hope it’ll be a nice day today
U: I want to eat grilled meat

The users also tend to use imprecise or awkward expressions, with an unspecified reference (“I said it to you”); give incorrect answers (B: “What is the difference between virginity and non-virginity?” — U: There’s no difference”), and ask either irrelevant questions (“Do you hear me?”) or too general ones (“What do you know?”), as well as echo questions (“Job?”). This all contributes to the overall lack in coherence and cohesion.

Additionally, the users prefer repetition to new attempts at evolving conversation:

U: It’s good.
B: A little.
U: You are wonderful.
B: Exactly. That’s being said...
U: You are really wonderful.
B: :) It’s hard.
U: I am thinking what to eat.

The above is a representative example of a conversation with the chatbot, illustrating that there is not much of a difference in the conversational styles of the chatbot and the user, whereby the user violates the cooperation principle on all fronts. Non-observance of manner and relevance is also demonstrated in scolding, ignoring attempts to show emotions, as well as nonsensical questions and responses (“Will you fire something?”, “That’s my line!”).

Dissatisfied with the course of a given chat (yet not blaming themselves), the users also signal prematurely that they wish to abort it (e.g. by typing “Bye” in the middle of a conversation), hence triggering a communication breakdown due to inability to rephrase, choose a different topic, make their contribution longer or, in general, avoid nonsensical turns in conversation.

This all demonstrates that chatbot developers cannot expect that the users will adhere to conversational maxims when dealing with a computer system. Thus, training on real human conversations can pose limitations when dealing with adjacency pairs appearing in actual dialogues with chatbots.

CONCLUSION

In this paper, we outlined the process of creating a rule-based chatbot system with a set of rules derived automatically from Twitter conversations. Our experience shows that Twitter can serve as a source of relatively long (10 or more lines) casual conversations between people, rich in informal language constructions.

The resulting system has a simple architecture, somewhat compensated with a large number of AIML rules (over 600,000 in the current implementation). Our experiments show that the system is able to engage in conversations with people, and keep track of the dialogue context to some extent. However, it appears that TF-IDF is not adequate enough to serve as a reliable relevance measurement in this task. It is clear that the chatbot’s responses are often irrelevant to the conversation at hand. It is also plausible that a single-line context, used by the bot, is not sufficient to keep track of the ongoing conversation. Furthermore, the selection of the most relevant dialogue lines according to TF-IDF measure produces predictable results. We are therefore planning to introduce random factors into the system.

Interestingly, since the users do not observe conversational rules on numerous occasions, those rules that definitely must not be violated should be specified both for the bot and for the user. It seems that the users tend to switch into the “evaluation mode” and “play” with the system to find out the chatbot’s response to their particular remarks, rather than take turns in a genuine dialogue. As our system is trained entirely on real conversations, it typically fails to find an adequate answer when faced with such challenges. Hence, the improved version of our chatbot should be able to recognise irrelevant input and attempt to steer the conversation back on track. Moreover, the users should either undergo training to learn to converse with a chatbot more successfully or not be informed about the fact that their conversation partner is non-human to avoid bias.

REFERENCES

- [1] B. AbuShawar and E. Atwell, “ALICE chatbot: trials and outputs,” *Computación y Sistemas*, vol. 19, no. 4, pp. 625–632, 2015.
- [2] R. Higashinaka *et al.*, “Towards an open-domain conversational system fully based on natural language processing,” in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 928–939.
- [3] T. Bickmore and J. Cassell, “Relational agents: a model and implementation of building user trust,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2001, pp. 396–403.
- [4] B. A. Shawar and E. Atwell, *A comparison between Alice and Elizabeth chatbot systems*: University of Leeds, School of Computing research report 2002. 19, 2002.

- [5] H. Yamaguchi and M. Mozgovoy, "Generating AIML Rules from Twitter Conversations," vol. Communication Papers of the 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 59–61, 2017.
- [6] F. Bessho, T. Harada, and Y. Kuniyoshi, "Dialog system using real-time crowdsourcing and twitter large-scale corpus," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2012, pp. 227–231.
- [7] Twitter Inc, *Twitter Streaming API*. Available: <https://dev.twitter.com/streaming/overview>.
- [8] R. Wallace, "The elements of AIML style," *Alice AI Foundation*, 2003.
- [9] J. Ramos and others, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 2003, pp. 133–142.
- [10] T. Kudo, *MeCab: Yet Another Part-of-Speech and Morphological Analyzer*. Available: <https://sourceforge.net/projects/mecab>.
- [11] ALICE A.I. Foundation, *Program AB*. Available: <https://code.google.com/archive/p/program-ab>.
- [12] C. Stratton, *PyAIML -- The Python AIML Interpreter*. Available: <https://github.com/creatorrr/pyAIML>.
- [13] C-W. Liu *et al.*, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation" *arXiv preprint arXiv:1603.08023*, 2016.
- [14] C. Chakrabarti and G. F. Luger, "A Framework for Simulating and Evaluating Artificial Chatter Bot Conversations," in *FLAIRS Conference*, 2013.
- [15] N. M. Radziwill and M. C. Benton, "Evaluating Quality of Chatbots and Intelligent Conversational Agents," *arXiv preprint arXiv:1704.04579*, 2017.
- [16] A. P. Saygin and I. Cicekli, "Pragmatics in human-computer conversations," *Journal of Pragmatics*, vol. 34, no. 3, pp. 227–258, 2002.
- [17] H. P. Grice, "Logic and conversation", in *Syntax and Semantics*, Vol. 3, *Speech Acts*, P. Cole, & J. L. Morgan, Eds. New York: Academic Press, 1975, pp. 41–58.
- [18] J. L. Austin, *How to Do Things with Words* Cambridge, MA: Harvard University Press, vol. 13, 1962.
- [19] J. R. Searle, *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press, 1969.
- [20] G. N. Leech, *Principles of pragmatics*. London: Longman, 1983.

6th International Workshop on Smart Energy Networks & Multi-Agent Systems

OUR energy supply infrastructure is in the middle of a transition from a conventional star-like energy supply topology with a manageable number of well-structured power plants towards a grid topology with a myriad of different generation units that are geographically widely distributed. Additionally, the increasing integration of volatile and intermittent renewable energy resources brings massive challenges to grid operations and its composition with respect to power system commitment, dispatching and reserve requirements.

The fact that renewable energy generation units will increase their share in the overall energy production, calls for technologies to be developed in the next decades to deal with the transition of the energy supply system and the distribution of renewable energy generation units. This includes technologies to integrate, handle and intelligently manage energy storage systems, grid load peak-shaving, smart supply system components, more efficient and intelligent coupling of heating with electrical power, heat storage, intelligent load shifting and balancing, to name only a few here.

All these have in common that the future power grid has to be intelligent, where generation and consumption units communicate or even negotiate their offer or their demand of energy through an ‘internet of energy’. Thus, to efficiently design and develop those distributed energy management systems is one of the key challenges to be solved to transform the energy supply system, addressing distributed coordination, as well as different forms of energy like electricity, heat, natural gas and other.

Information and communication technologies are the key enablers of such envisioned systems, where especially the agent-paradigm provides an excellent modelling approach for the distributed character of energy systems. Although significant efforts and investments have already been allocated into the development of smart grids, there are, however, still significant research challenges to be addressed before the promised efficiencies or visions can be realised. This includes distributed, collaborative, autonomous and intelligent software solutions for simulation, monitoring, control and optimization of smart energy networks and interactions between them.

TOPICS

The SEN-MAS’18 Workshop aims at providing a forum for presenting and discussing recent advances and experiences in building and using multi-agent systems for modelling, simulation and management of smart energy networks. In particular, it includes (but is not limited to) the following topics of interest:

- Experiences of Smart Grid implementations by using MAS
- Applications of Smart Grid technologies
- Distributed energy management of distributed generation and storage based on MAS
- Examples of design patterns for MAS in distributed energy management systems
- Microgrids, Islands Power Systems
- Real time control of energy networks
- Distributed planning process for energy networks by using MAS
- Self-configuring or self-healing energy systems
- Load modelling and control with MAS
- Simulations of Smart Energy Networks
- Software Tools for Smart Energy Networks
- Energy Storage
- Electrical Vehicles
- Charge scheduling for electric vehicles (and fleets) based on MAS
- Interactions and exchange between networks for electricity, gas and heat
- Stability in Energy Networks
- Distributed Optimization in Energy Networks
- Safety and security issues for MAS in Smart Grids

TUTORIAL

Beside the scientific exchange, the event will provide a practical tutorial for building so called Energy Agents. Based on the open-source framework Agent.GUI, the JADE agent platform and the experiences in the project Agent.HyGrid, the tutorial will guide you through the development process that enables you to build agents that can be installed and executed beside distributed energy systems.

EVENT CHAIRS

- **Brehm, Robert**, University of Southern Denmark, Denmark
- **Derksen, Christian**, University Duisburg-Essen, Germany

PROGRAM COMMITTEE

- **Bilal, Bilal**
- **Bremer, Joerg**, joerg.bremer@uni-oldenburg.de, Germany
- **Fortino, Giancarlo**, Università della Calabria
- **Hildmann, Hanno**, Universidad Carlos III de Madrid (UC3M), Spain

- **Karnouskos, Stamatis**, SAP, Germany
- **Klusch, Matthias**, German Research Center for Artificial Intelligence, DFKI, Germany
- **Loose, Nils**
- **Moench, Lars**, FernUniversität Hagen, Germany
- **Nieße, Astrid**, Leibniz Universität Hannover, Germany
- **Paprzycki, Marcin**, Systems Research Institute Polish

Academy of Sciences, Poland

- **Redder, Mareike**
- **Sonnenschein, Michael**, Professor (retired) at the University of Oldenburg, Germany
- **Sudeikat, Jan**, Hamburg Energie GmbH, Germany
- **Vale, Zita**

Design of models for the tokenization of electric power industry basing on the blockchain technology

Andrew Varnavskiy
Financial University under the
Government of the RF,
Leningradsky Prospekt 49,
Moscow, Russia
Email: AVVarnavskiy@fa.ru

Ulia Gruzina
Financial University under the
Government of the RF
Leningradsky Prospekt 49,
Moscow, Russia
Email: ymgruzina@fa.ru

Artur Rot
Wroclaw University
of Economics,
ul. Komandorska 118/120,
53-345 Wroclaw, Poland
Email: artur.rot@ue.wroc.pl

Vladislav Trubnikov
Financial University under the
Government of the RF
Leningradsky Prospekt 49,
Moscow, Russia
Email: vladtrubnikov95@gmail.com

Anastasia Buryakova
Financial University under the
Government of the RF
Leningradsky Prospekt 49,
Moscow, Russia
Email: AOBuryakova@fa.ru

Ekaterina Sebechenko
Financial University under the
Government of the RF
Leningradsky Prospekt 49,
Moscow, Russia
Email: EVSebechenko@fa.ru

Abstract—The problem of implementing modern technologies into the electric power industry is quite relevant in the world. The article considers the models of decentralized platforms providing services for energy distribution and trade, their main advantages and disadvantages. The basic principles of tokenization were developed, which allow optimizing of the energy systems and concentration of the crowd funding process for the construction of new generation facilities.

I. INTRODUCTION

Nowadays, an entire "ecosystem" of companies has been formed around the concept of a distributed ledger, which build their work using internal accounting units. Thanks to the technology of the distributed registry, it becomes possible to optimize the generating and trading process of power capacities, which ensures a high demand from companies from all over the world.

The aim of the work is to identify the most optimal model for the operation of the decentralized energy platform to increase the efficiency of the generation and distribution of electric power.

In order to systematize the obtained data, we use methods of retrospective, current and prospective analysis and synthesis of theoretical and practical material.

The peculiarity of blockchain technology lies in the fact that the network ledger is updated every time as a transaction occurs. Suppliers and consumers determine the variables of a given transaction by allocating the specific parameters, address of sender, sum and amount of energy, as well as other necessary information, after which the variables are

combined with detailed information on other transactions conducted during the same period of time to create a new block data. If any information relating to a particular transaction is subsequently changed as a result of unauthorized interference or due to data transmission errors, the algorithm will report an error.

There are allocated characteristics among the merits of the blockchain technology: reduction of transaction costs; ensuring maximum transparency of operations; reducing the overall complexity of transactions; storage of accurate data on all transactions performed. The shortcomings of the technology are mainly due to the lack of the necessary infrastructure and regulatory component: an undefined status of tokens (special kind of virtual currency that represent an asset or utility) – accounting units of the system, which are used, among other things, to purchase goods; lack of a regulatory framework for dispute resolution in some countries of the world; possible technical problems and failures in the initial stages of the technology introduction; complete loss of control over the tokens (system currency) in circulation in the case of cyberattacks. From the point of view of the electricity market, it should be noted that all energy transactions can be accurately recorded for certain suppliers and consumers. Controlling the distributed and consumed energy will inevitably lead to its most effective use, and also enable producers to plan generation in advance. These features become more sustainable under the system of Smart Grid operations. The platform of blockchain-options trading can provide a necessary level of automation using the smart contracts, which will reduce the costs and helps to directly connect consumers and suppliers of energy.

The material has been prepared with the results of studies carried out at the expense of funds provided under the grant of the Bank Santander.

The platform also has its own investment foundation for suppliers, which goes as an addition, but can help form a system of trust and discounts, especially for large buyers.

Despite having some automated functions, our platform cannot be classified as a Multi-agent system, because it still has to have a lot administrative control over the course of its work.

II. RELATED WORK

During the study, a special attention was paid to the documentation of already existing energy projects: Greeneum [1], Suncontract [2], Grid + [3], Impact PPA [3], Power Ledger [4], Energo Labs [5], Exergy project [6], Enerchain [7], WePower [8]. A thorough analysis of the energy tokenization (the process of converting rights to real world assets into a digital token on a blockchain) proposed by the projects made it possible to identify the main principles on which the energy trade platform should be built. The provisions of the EY [9], PWC [10], Deloitte [11] reports allowed to look at the possibilities of using blockchain technology in the electric power industry. Confirmation of ownership and trading in emission allowances for pollutants are truly promising and original ideas.

Moreover, the experience of some foreign researchers was studied and analyzed. For example, Basden J., Cottrell M. in the work "How utilities are using blockchain to modernize the grid" [12] surely indicate that the blockchain is a reliable, inexpensive way to conduct and control transactions without a central generation unit of power. According to the authors, the blockchain will promote the development of renewable energy microgrid. Alternatively, Thomas Morstyn, Niall Farrell, Sarah J. Darby and Malcolm D. McCulloch in their article "Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants" [13] drew attention to the prospects of introducing intelligent solutions and energy management systems to ensure universal and safe access to energy. Technical problems and recent achievements, identified in the work "Blockchain Challenges and Opportunities" [14] Zheng Z. et al. also found their reflection in our article.

During the analysis of the market we used the experience of Russian scientists. The following works were considered: Bogdanova ED, Valieva LG "Cryptocurrency and energy" [15], Veselov FV, Khokhlov A. "Internet of Energy: how distributed energy will affect safety, prices for electricity and ecology" [16], Lyalkov IM "Management of financial risks in companies of the fuel and energy complex" [17], etc.

III. ANALYSIS OF EXISTING BLOCKCHAIN MODELS IN THE ENERGY SECTOR

The tokens issued during the ICO can perform a variety of functions – payment for goods, works, services provided by the company; various discounts and bonuses; means of

payment; system access keys; confirmation of ownership of an asset or a normal means of attracting financing, etc. For example, a token is the accounting unit of a system whose properties depend solely on how the community will use it. Tokens can be secured with property, have a constant inflation rate, the company can also provide some monetary policy. The number of tokens produced by projects is also not limited, which allows creating different models that meet the requirements of the platform. Different models of tokenization of a number of energy projects were analyzed in the course of the study.

WePower is a platform that allows producers of renewable sources of energy to tokenize and sell generated power. Joining the platform, the manufacturer receives the number of tokens, and get equally supplied energy. 1 kWh is equal to 1 token. Each energy supplier, connected to the platform, organizes an auction for the sale of tokenized energy. Consumers who own WePower tokens have priority access to auctions. The amount of energy they acquire will depend on the number of WePower tokens that they have. The auction starts with the lowest, base price (1 kWh = 1 WPR). Holders of WPR tokens have extended access for the first 48 hours, after which the non-sold energy is offered to all participants of the platform. Current and historical energy prices are fixed and displayed in the application, which allows participants to focus on the market situation.

An important point is that WePower energy suppliers get the right to sell tokens to consumers without real deliveries at the time of the transaction, because a token is in form of obligation of the supplier to provide a certain amount of energy in the future. At the same time, suppliers with generation facilities and connected to the platform can receive a fee in WePower tokens or fiat money at the time of the transaction by giving energy.

In addition to the indicated advantages, the above models also have significant drawbacks. Thus, speculation on the exchange rate of coins of platforms is not ruled out. The offer of the token will depend wholly and entirely on the number of users providing and purchasing energy. The more suppliers, the more coins and the lower the price in conditions of weak demand. However, in the case of a very limited supply and high demand, attempts of fraud in the market cannot be brushed away, when profits will be used by suppliers for speculative purposes.

In order to consider more complex models, it is necessary to have a general idea of the operation of the power grid. The existing power system assumes the following stages of the process of bringing energy from the producer to the recipients: generation, transmission, distribution and retail. Generation, subsequently, is the production of energy by large producers of all types of electricity: hydrocarbon, nuclear, hydroelectric, etc. Recently, wind and solar farms has started to gain popularity. Transmission is the movement of electricity over long distances, usually "from generators" to "distributors". Distribution – the process of electricity

transmission from high voltage networks to end users. "Distributors" work with low-voltage electric lines, which are connected to households or enterprises. The last stage is the stage of trade and will affect the changes caused by decentralization.

Here you can use a purely market-based instrument – a forward contract. Forward – an agreement by which one party (in our case the supplier) undertakes to transfer the goods (energy) to another party (buyer) in a specific contract, and the buyer agrees to accept and pay this asset on time. In the case of the energy market, forward contracts would be more reliable, since the supplier would receive a guarantee that all the energy produced will be realized, and the consumer would not have doubts about the sudden rise in prices. It would be possible to prevent part of the attempt to avoid speculation. However, such arguments will not always be appropriate.

It is necessary to pay special attention to the fact that a strong fluctuation of prices often occurs when tokens are not provided with any real goods. Confirmation of this thesis is the first model with one token, where a part of the coins has not actually been backed up by energy from the moment of entering the stock exchange. Tokens passed into the category of purely speculative instruments. The model with two tokens is specifically designed to eliminate this drawback by using various monetization mechanisms.

Power Ledger. The market flexibility of the ecosystem is designed to support the model with tokens performing the functions reflected in Table I.

The meaning of tokenization lies not only in assigning certain information to the accounting units to rationalize the distribution of energy, but also in the formation of fair prices. Some number of provisions are fixed in the process of supplying energy: volume, source (renewable / non-renewable), geographical location of generating capacities, etc. These criteria are forming the basis of the price. That is, the token in this case is not a mean of calculation, but only a marker that helps the community to determine the price. Thanks to the use of smart contracts, it is possible to unite suppliers and consumers, minimizing the costs of both parties. Specificity of the tokenization model is described in the accompanying document "Exergy Business".

Based on the description of technology, we can distinguish the following functions of the internal accounting unit:

- identification of each physical operation (creation, transmission, storage, consumption of energy);
- ensuring the ownership of the main product of the project – energy;
- ensuring confidentiality and security of the main transactions;
- differentiation of generation types by the location of generating units and other significant and important characteristics.

TABLE I.
"POWER LEDGER" TOKENIZATION MODEL

POWR	Sparkz
<ol style="list-style-type: none"> 1. Traded on the stock exchange; 2. Required to access the platform; 3. Guarantees ownership of a renewable energy asset; 4. Charged to suppliers of renewable energy as a reward for loyalty; 	<ol style="list-style-type: none"> 1. Issued in return for POWR through Smart Bond; 2. Unit of energy cost in local markets around the world; 3. Can be exchanged for cash;

The description of the model from a technological point of view suggests that Exergy is going to ICO offering two tokens. One will be used only for internal purposes, and the second will be used to attract additional funding and new participants.

Of all the projects reviewed, five ICOs have already taken place. The Grid + energy trading platform following the Pre-ICO results, which ended in July 2017, managed to attract \$ 45.076 million (Table II). Investments attracted by the WePower platform amounted to 40 million US dollars. The planned amount was collected in one day, ICO was stopped ahead of schedule. The Power Ledger project in September 2017 attracted \$ 13.23 million, Suncontract – \$ 2.5 million, EnergoLabs – \$ 2.2 million. So, the most successful strategies for attracting investments are Grid + and WePower solutions. Prior to joining the ICO, the companies already had agreements with energy companies, could present the tested technological solutions and conducted an active marketing policy.

Thus, analyzing the efficiency of blockchain technology on the market sites, it is worth taking into account both the direct advantages of applying the technology for energy exchange and the business plans of new companies, because the attraction of financing was also due to the issuance of digital accounting units of the system.

TABLE II.
THE VOLUME OF FINANCING ATTRACTED BY THE PROJECTS
IN THE SECTOR OF POWER SUPPLY

Project	Stage (ICO/Pre-ICO)	Date of closing ICO	The volume of attracted financing, thousands USD
Suncontract	Pre-ICO	28.07.2017	2 501
Grid+	Pre-ICO	30.11.2017	45 076
Power Ledger	ICO	08.09.2017	13 232
EnergoLabs	ICO	15.08.2017	2 228
WePower	ICO	02.02.2018	40 000

The general model of the considered energy decentralized platforms is indicated in Fig. 1.

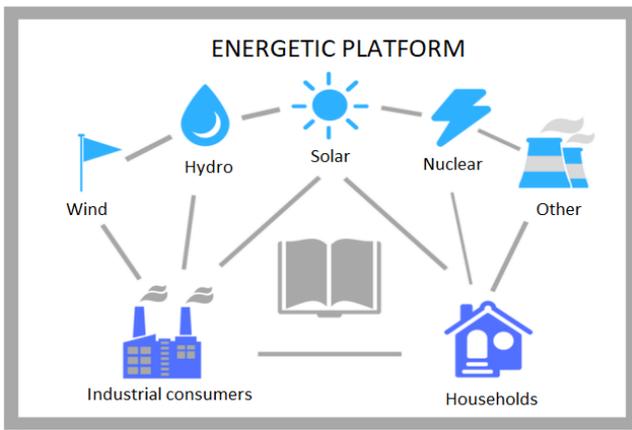


Fig. 1 Model of the decentralized energy area

Despite the advantages of a platform based on blockchain technology, it is still difficult to talk about the effectiveness of its work in terms of comparing costs and financial results, since none of the projects has been fully implemented. Nevertheless, theoretically the benefits to suppliers and consumers are undeniable. Elimination of intermediaries and introduction of P2P transactions will inevitably lead to a reduction in energy prices under the conditions of high competition. It is worth noting the simplicity of the issue and sale of coins in the market as an investment tool. In ICO, all interested users can participate in the project development. Consumers will receive tokens, which can later be exchanged for electricity, investors after placing the coin on the exchange will be able to sell it at a market price. With a competent approach to the organization of a trading power market, it becomes possible to substantially optimize the work of the industry as a whole.

IV. THE MAIN CHARACTERISTICS OF THE ENERGY PLATFORM MODEL

An integrated tokenization scheme was developed without the use of derivative financial instruments based on the analysis of models with two interconnected tokens, presented on the market (Fig. 2).

This scheme provides for six steps:

Step 1

At the initial stage, consumers / investors purchase an external token No. 1 equal to a certain amount of energy X kWh per ICO. The received money is sent to the project fund for further development of the platform, team reward, marketing, etc. At the same time, it is necessary to envisage the accrual of an internal token No. 2 to the accounts of consumers in exchange for a fiat currency. This need arises as a result of the fact that the number of external tokens will decrease with time due to them being burned out after the conclusion of transactions for the supply of energy.

The suppliers of energy receive their portion of tokens by signing agreement with the administration of the platform.

Later on they will appear on the consumers` wallets and exchange and eventually burned out.

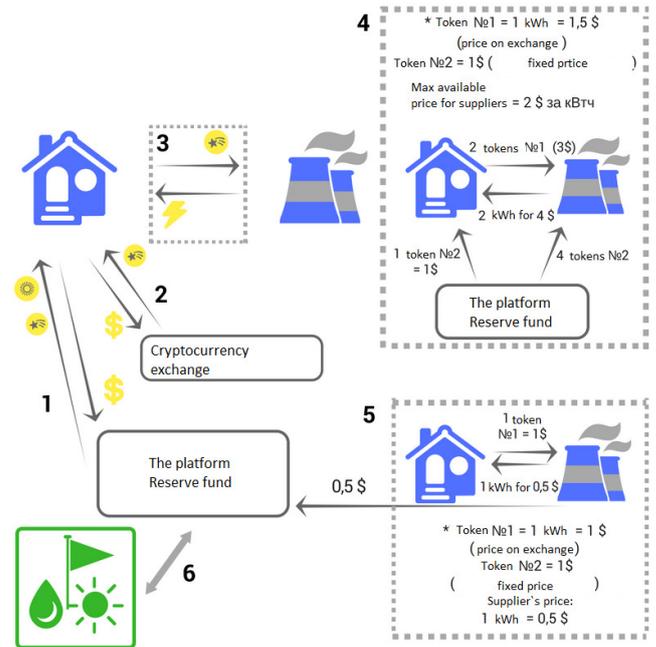


Fig. 2 Energy tokenization scheme

Step 2

Exit to the exchange is provided only for the external accounting unit – token No. 1. It is assumed that the price of the coin will grow due to an increase in demand and a constant reduction in their quantity in circulation. Thanks to this, energy suppliers will be able to receive compensation that allows them to compensate for energy generation costs.

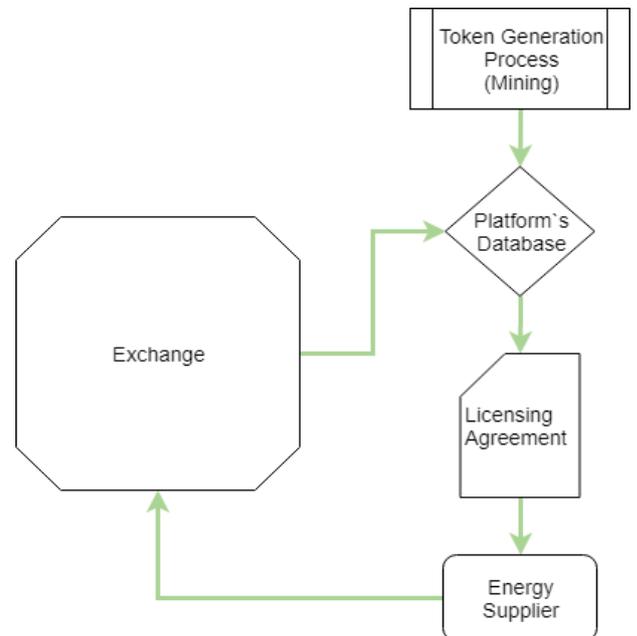


Fig.3 Token Generation process

Step 3

Before users completely switch to transactions using token No. 2, consumers will use tokens №1 to purchase energy, but it is clear that the exchange in internal tokens is more convenient. Thus, the token №1 turns into a purely market instrument, backed by the right to purchase energy. In the market there can be two situations: 1 - the exchange price of the token No. 1 will be higher than the maximum possible price of supply and 2 - the exchange price will be lower than the prices set by the supplier.

*Step 4

If the price of the token No. 1 turns out to be lower than the prices set by the supplier, the consumer will receive compensated overpaid funds in tokens № 2. For example, if the token №1 at the exchange price of \$ 1 will give the right to purchase 1 kWh, and the supplier will be able to supply the required amount of energy only at a price of \$ 2, the consumer will be paid compensation in the amount of 1 US dollar equal to 1 internal token. But this situation is not so feasible. Presumably, the price of the token at the time of placement should be several percent higher than the market price, and further growth in the exchange rate should be supported not only by burning, but by an increase in demand for a constantly rising price of an asset.

*Step 5

The situation in which the price of the token No. 1 exceeds the price of suppliers, will be taken as the base one. Consumers will be able to use the guaranteed right to supply energy or refund, and suppliers will sell products at affordable prices. The greater the difference between the price of supply and demand, the greater the percentage of remuneration received by suppliers and the greater the amount of funds used to replenish the project's reserve fund.

However, it should be noted that such mechanism will take place until all tokens № 1 are eliminated. The technical token No. 2, which the whole platform should go over in time, will be equivalent to the national currency, which means that the system will operate according to a purely market mechanism, where demand balances the offer. In the future, new placements can be conducted to attract additional funding.

Step 6

The formation of the reserve fund will not only support the operation of the platform, but also reinvest the profits received for the construction of new generation facilities. Financing the construction of new stations will help increase the energy suppliers connected to the platform. The use of smart contracts to conclude transactions and the growth of the energy potential of the entire system will lead to the creation of the best terms of trade at low prices.

Thus, the model with two tokens was recognized as the most effective on the basis of the analysis of decentralized companies which provide services for energy distribution

and trade. It allows to distinguish between trading on the exchange (coin/national currency) and intra-system trade (coin/energy). This approach also excludes the possibility of speculation, because each platform user can exchange a token only for energy. At the same time, minimization of risks occurs due to the formation of a reserve fund, which can be used in the emergency: a sharp increase in the costs of production of a supplier or a sudden drop in the solvency of consumers. The difference between the price of supply and demand will be offset by the resources accumulated by the system.

V. CONCLUSION

Thus, the following perspectives of the blockchain technology in the supply of electric power were identified the analysis of the Russian electricity market:

1. The optimization of the functioning of power systems.
2. Development of renewable energy sources by attracting additional financing through the ICO for the construction of energy generation facilities.

The analysis of the existing energy blockchain projects allowed us to identify the two most successful investment strategies – Grid + and WePower. The key factors were the signed agreements with energy companies, the availability of tested technological solutions and active marketing policy. However, it is still difficult to estimate the effectiveness of the projects reviewed in terms of comparing costs and financial results, since they are not fully implemented at the moment. Nevertheless, theoretically the benefits to suppliers and consumers are undeniable. It is established that the liquidation of intermediaries and introduction of P2P transactions will inevitably lead to a reduction in energy prices under the conditions of high competition.

REFERENCES

- [1] *Official site of Greeneum project*, White paper, <https://www.greeneum.net/greeneum-whitepaper/>
- [2] *Official site of Suncontract project*, White paper, <https://suncontract.org/tokensale/res/whitepaper.pdf>
- [3] *Official site of Grid+ project*, White paper, <https://gridplus.io/assets/Gridwhitepaper.pdf>
- [4] *Official site Impact PPA project*, White paper, https://www.impactppa.com/wp-content/uploads/2018/03/ImpactPPA_WP_v1.2WEB.pdf
- [5] *Official site of Power Ledger project*, White paper, <https://powerledger.io/media/Power-Ledger-Whitepaper-v8.pdf>
- [6] *Official site of Energo Labs project*, White paper, <http://www.energolabs.com/>
- [7] *Description of the Enerchain project*, <https://enerchain.ponton.de/>
- [8] *Official site WePower project*, White paper, https://drive.google.com/file/d/0B_OW_EddXO5RWWFVQjJGZXpQT3c/view
- [9] EY, *Overview of the Electricity Industry* <http://ru.investinrussia.com/data/file/EY-power-market-russia-2018.pdf> [In Russian].
- [10] PwC, *Use cases for blockchain technology in Energy & Commodity trading – 2017*, <https://www.pwc.com/gx/en/industries/assets/blockchain-technology-in-energy.pdf> [Accessed 08.03.2018].
- [11] Deloitte, *Blockchain applications in energy trading – 2016*, <https://www2.deloitte.com/global/en/pages/energy-and->

- resources/articles/role-of-blockchain-in-the-energy-and-resources-industry.html [Accessed 28.04. 2018].
- [12] J. Basden, M. Cottrell, *How utilities are using blockchain to modernize the grid*, Harvard Business Review, 2017.
- [13] T. Morstyn, et al., *Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants*, Nature Energy, T. 3, No. 2. – C. 94, 2018.
- [14] Z. Zheng et al., *Blockchain challenges and opportunities: A survey*, Work Pap., 2016.
- [15] E.D. Bogdanova, L.G. Valieva, “Cryptocurrency and energy. Problems, prospects and tendencies of innovative science development”, *Collected articles of international academic and research conference – 2017*, pp. 63-67 [In Russian].
- [16] F.V. Veselov, A. Khokhlov, *Internet of Energy: how distributed energy will affect safety, prices for electricity and ecology*, Russian version of the information resource Forbes, section "Business", October 18, 2017 [In Russian].
- [17] I.M. Lyalkov, “Management of financial risks in companies of the fuel and energy complex”, Vestnik of the Russian Economic University. G.V. Plekhanov. Introduction. The way to science No. 1 (17), 2017, pp. 93-100 [In Russian].
- [18] *Official site of the system operator of the Unified Energy Network of Russia. Ensuring the long-term development of the EEC*, <http://so-ups.ru/index.php?id=future>
- [19] A.A. Burdin, *Prerequisites for the Sustainable Development of the Electric Power Industry of Tatarstan*, Synergy of Sciences No. 10., 2017, pp. 41-45, [In Russian].
- [20] *Exergy project Technical Whitepaper*, <https://exergy.energy/wp-content/uploads/2017/12/Exergy-Whitepaper-v8.pdf>
- [21] MarketLab, *Metallgesellschaft* <https://market-lab.org/kejs-metallgesellschaft>
- [22] U.S. Commodity futures trading commission, <https://www.cftc.gov/About/MissionResponsibilities/index.htm/>
- [23] A.M. Rudkevich, M.C. Caramanis, E.A. Goldis, L. Xiaoguang, P.A. Ruiz, “Tabors R.D. Financial Transmission Rights in Changing Power Networks”, *Proceedings of Hawaii International Conference on System Sciences*, No. 49, 2016, pp. 2326-2334.
- [24] Arstechnica, *Bitcoin's insane energy consumption, explained*, <https://arstechnica.com/tech-policy/2017/12/bitcoins-insane-energy-consumption-explained/>
- [25] M. Kindy, A. Divine, *A Blockchain Reputation System For Determining Good Market Actors*, <https://medium.com/topl-blog/divine-a-blockchain-reputation-system-for-determining-good-market-actors-7c47a0308ae8/>
- [26] T. Dickerson, M. Herlihy, P. Gazzillo, E. Koskinen, “Adding Concurrency to Smart Contracts, Computer Science” *Distributed, Parallel, and Cluster Computing*, Cornell University Library, <https://arxiv.org/abs/1702.04467>
- [27] M.J. Mamontova, “Blockchain and opportunities of its implementation in energy, Information technology in science, management, social sphere and medicine: collected scientific proceedings of IV International scientific conference, December 5-8 2017, Tomsk, pp. 417-419 [In Russian].

11th International Workshop on Computational Optimization

MANY real world problems arising in engineering, economics, medicine and other domains can be formulated as optimization tasks. These problems are frequently characterized by non-convex, non-differentiable, discontinuous, noisy or dynamic objective functions and constraints which ask for adequate computational methods.

The aim of this workshop is to stimulate the communication between researchers working on different fields of optimization and practitioners who need reliable and efficient computational optimization methods.

TOPICS

The list of topics includes, but is not limited to:

- combinatorial and continuous global optimization
- unconstrained and constrained optimization
- multiobjective and robust optimization
- optimization in dynamic and/or noisy environments
- optimization on graphs
- large-scale optimization, in parallel and distributed computational environments
- meta-heuristics for optimization, nature-inspired approaches and any other derivative-free methods
- exact/heuristic hybrid methods, involving natural computing techniques and other global and local optimization methods

The applications of interest are included in the list below, but are not limited to:

- classical operational research problems (knapsack, traveling salesman, etc)
- computational biology and distance geometry
- data mining and knowledge discovery
- human motion simulations; crowd simulations
- industrial applications
- optimization in statistics, econometrics, finance, physics, chemistry, biology, medicine, and engineering.

BEST PAPER AWARD

The best WCO'18 paper will be awarded during the social dinner of FedCSIS 2018.

The best paper will be selected by WCO'18 co-Chairs by taking into consideration the scores suggested by the reviewers, as well as the quality of the given oral presentation.

EVENT CHAIRS

- **Fidanova, Stefka**, Bulgarian Academy of Sciences, Bulgaria
- **Mucherino, Antonio**, INRIA, France
- **Zaharie, Daniela**, West University of Timisoara, Romania

PROGRAM COMMITTEE

- **Abud, Germano**, Universidade Federal de Uberlândia, Brazil
- **Bonates, Tibérius**, Universidade Federal do Ceará, Brazil
- **Breaban, Mihaela**
- **Chira, Camelia**, Technical University of Cluj-Napoca, Romania
- **Gruber, Aritanan**
- **Hosobe, Hiroshi**, Hosei University, Japan
- **Iiduka, Hideaki**, Kyushu Institute of Technology, Japan
- **Lavor, Carlile**, IMECC-UNICAMP, Brazil
- **Micota, Flavia**, West University of Timisora, Romania
- **Muscalagiu, Ionel**, Politehnica University Timisoara, Romania
- **Pintea, Camelia**, Tehnical University Cluj-Napoca, Romania
- **Siarry, Patrick**, Universite Paris XII Val de Marne, France
- **Stefanov, Stefan**, South-West University "Neofit Rilski, Bulgaria
- **Stoean, Catalin**, University of Craiova, Romania
- **Stuetzle, Thomas**, Université Libre de Bruxelles (ULB), Belgium
- **Zilinskas, Antanas**, Vilnius University, Lithuania

"Passeport Vacances": an assignment problem with cost balancing

Corentin Beffa, Sacha Varone

University of Applied Sciences Western Switzerland (HES-SO),

HEG Genève, Switzerland

Rue de la Tambourine 17, 1227 Carouge, Switzerland

Email: {corentin.beffa, sachavarone}@hesge.ch

Abstract—Passeport Vacances is an offer for school-aged children to discover a set of activities during holidays. For more than 30 years, it has been an established social function in several countries, including Germany and Switzerland. Proposed activities might occur several times during the Passeport Vacances. The assignment of activities to children is computed in order to maximize the children's preferences, as well as to balance each child's incurred cost, toward an equity goal. There are several sets of constraints associated with the assignment problem: no overlapping activities assigned to the same child, minimal and maximal ages per activity, minimum number of children for opening an activity, maximal size of a group for each activity, no similar activities assigned to the same child, no already assigned 'lifetime'-activity per child, and at most one activity per period and per child. We propose a binary linear programming model that describes the assignment problem, report CPU computation issues regarding the model implementation, and report numerical results based on a state-of-the-art MIP solver. Tests were conducted with real data from the 2016 edition of Passeport Vacances in Morges.

is divided into periods, most often days or half-days. Before PV actually takes place, children are asked to select a few activities per period, often 4 selections ranked by preference. After closing this selection phase, the assignment process can begin.

Many areas are interested in load balancing: for example, to balance the workloads of teaching assistants [1], students [2], or professors [3] or for socioeconomic variation between schools [4]. This balancing is often done by minimizing the deviation to the mean value or, as proposed by Domenech and Lusa [5] and De La Torre, Lusa and Mateo [3], combined to the maximum relative deviation. However, the different measures of deviation are not linear and need to be adapted. Ünal and Uysal [2] proposed a linearisation of the 4 norms L_0 , L_1 , L_{inf} , and L_{max} by adding variables. The balancing objective is often one of several goals of the problem. Different ways to combine multiple goals have been proposed in the literature. In the purpose of assigning students to projects, Pan, Chu, Han, Guangyue, and Huang [6] proposed a goal programming model. Another well-known method consists of making a mixed integer linear programming (MILP) model as a weighted sum of the different objectives. The weights can be adapted to privilege one or another goal, but many studies have focused on finding a Pareto optimal solution: a solution where all objectives could not be improved without deteriorating the others. A small literature review and a method to find the Pareto front was proposed by Kim and De Weck [7]. In order to propose a convenient and powerful model, we chose to use a weighted sum of the different goals with fixed weights.

This paper reports our mathematical model, implemented in the Julia language and solved with a Mixed Integer Programming solver. State-of-the-art solvers are able to pre-compute the MIP model before launching the Simplex algorithm, so that redundant constraints are dropped, and also, as a consequence, some variables are fixed. Although this generally saves quite a lot of computation time, it is sometimes recommended to avoid part of this step by changing a straightforward implementation with a more data-focused model. The goal is to reduce the time needed to build the model, which could seriously increase the total CPU computing time. We explain in this report our verification and the way we handled the model construction. We then report experimental results and give some insight on the resulting computing time.

I. INTRODUCTION

PASSEPORT Vacances is an offer for school-aged children to discover a set of activities during holidays. For more than 30 years, it has been an established social function in several countries, including Germany and Switzerland. Proposed activities might occur several times during the Passeport Vacances. Some weeks before the start of Passeport Vacances, children are asked to choose at most four activities per day and to give a ranking to the selected activities for each day, from 1 for the most attractive one, to 4 for the least attractive one. Each child receives a personal identifier and has to give his birthdate, as well as other useful information such as phone number, address, etc. The assignment of children to activities is computed in order to maximize the preferences specified by the children, as well as to balance the incurred cost by each child, toward an equity goal. There are several sets of constraints associated with the assignment problem: no overlapping activities assigned to the same child, minimal and maximal ages per activity, minimum number of children for opening an activity, maximal size of a group for each activity, no similar activities assigned to the same child, no already assigned 'lifetime'- activity per child, and at most one activity per period and per child. During a given horizon of time, generally between 5 and 14 days, Passeport Vacances offers a set of activities to children during holidays. This holiday time

TABLE I
ACTIVITY

idactivity	unique identifier
nboccurrence	number of occurrences during the considered horizon
pricefixed	fixed price
pricechild	price per child
life	binary indicator about the status of ‘lifetime’
minchild	minimum number of children per occurrence to open the activity
maxchild	maximal number of children per occurrence
minage	minimal age to perform the activity
maxage	maximal age to perform the activity
similarity	identifier of similar activities

TABLE II
OCCURRENCE

idoccurrence	unique identifier
idactivity	associated activity
occurrencebegin	begin date and time of the occurrence
occurrenceend	end date and time of the occurrence
inactive	binary indicator variable
next	next occurrence in case of multi-occurrence activity
previous	previous occurrence in case of multi-occurrence activity

TABLE III
CHILD

idchild	unique identifier
birthdate	birthdate

II. PROBLEM DESCRIPTION

A. Data

Data was stored in a PostgreSQL database. It is therefore presented as a set of tables with their fields.

Table I specifies the activities proposed during the considered PV horizon.

- Field ‘fixedprice’ is the fixed cost that PV has to pay for each occurrence of the activity that occurs, whereas ‘pricechild’ is the variable cost per child that PV has to pay. Those costs can be considered in the following way: a balanced cost between children of the same age category is appreciated, so that a fair assignment of children to activities can be done.
- The ‘life’ field indicates if an activity has to be considered as a lifetime activity, which means that if a child has already been assigned to that activity in some past PV, he can no longer be assigned to this activity. Its value is TRUE for a lifetime activity, and FALSE otherwise.
- Field ‘minchild’ indicates the minimum number of children to open an occurrence of an activity. In other words, if there are not enough children for a specific occurrence, then this occurrence is cancelled. Field ‘maxchild’ restricts the size of the group for an occurrence of the activity.
- Fields “minage” and “maxage” are respectively limitations on the minimal age and maximal age for doing an activity.
- Field ‘similarity’ indicates similar activities. For example, activity A = visit to the zoo Alpha and activity B = visit to the zoo Beta are considered similar and therefore belong to the same similarity group. The goal of this field is to avoid similar activities being assigned to a same child. Modalities of the ‘similarity’ field are natural numbers.

Table II contains the occurrences of the activities. The fields ‘next’ and ‘previous’ get the value of ‘idoccurrence’ for activities that only need one occurrence to be done. For activities requiring consecutive occurrences, like a several days internship within a company, the ‘next’ field refers to the next occurrence, unless it is the last one; in this case, it contains the same value as ‘idoccurrence’. The ‘previous’

field is similarly defined. Cancelled occurrences are referred to with the ‘inactive’ field, whose value is TRUE for inactive occurrences and FALSE otherwise.

Table III contains the characteristics of children. Note that all personal information, such as first name, family name, phone number, etc., were not provided since they are useless for the optimization process. The day of each birth date has been modified to the first day of the month in order to anonymize the data and their IDs have been changed.

Table IV represents k-tuples of children, indicating groups of children willing to participate together, for each of the assigned activities. This is generally useful in case of children belonging to a same family, or friends willing to share activities.

Note that this not only refers to binomes, but also *k*-nomses (i.e. several children with the same set of assigned occurrences) can be constituted.

Table V indicates ranked preferences given by the children. Each child ranks at most *k* ($k=4$ in this data-set) activities per day of the PV, 1 being the preferred activity, up to *k* being the *k*th preferred activity.

Table VI expresses already-assigned ‘lifetime’ activities to children. This is a history of past PV assignments. A child

TABLE IV
KNAME

idchild1	first child
idchild2	second child

TABLE V
PREFERENCE

idpreference	unique identifier
idchild	child identifier
idoccurrence	occurrence identifier
choice	choice rank of the occurrence

TABLE VI
LIFETIME

idchild	child identificator
idactivity	activity identificator
idlifetime	unique identificator
idchild	child identificator
idactivity	activity identificator
idpreference	preference identificator

TABLE VII
PERIOD

idperiod	unique identificator
periodbegin	beginning of the period
periodend	end of the period
maxassigned	maximum number of occurrences for this period

who already received a specific lifetime activity in the past PV cannot again obtain this activity.

Table VII represents periods on which restrictions might apply. This means that during a specific period of time 'idperiod', beginning at 'periodbegin' and ending at 'periodend', a maximum of 'maxassigned' occurrences can be assigned to the same child.

B. Data preparation

In order to facilitate the model generation, we modified raw data by adding computed tables, adding some columns to existing tables, or reducing unnecessary information.

Passeport Vacances partitions the time horizon into periods, which are most often days but sometimes half-days. Children are then asked to fill a formula to express their preferences about activities. They must do this for each period.

We computed the incompatibility graph of preferences, which represents overlapping occurrences that appear in the selected activities by a same child.

Based on these inequalities, we created a new table called *incompatible* that contains all incompatible preferences, one for each edge of the incompatibility graph.

TABLE VIII
INCOMPATIBLE

idincompatible	unique identificator
idpref1	first preference identificator
idpref2	second preference identificator

III. MATHEMATICAL MODEL

We use the notation proposed in Table IX.

There is a set O of occurrences, because an activity can be organised several times during the whole Passeport Vacances period. Therefore, several occurrences might correspond to a same activity. A preference $p \in P$ refers to a child's choice for an occurrence. The sets A , T , C , I , L , MA_t , and C_p are

TABLE IX
NOTATIONS

P	Set of preferences
A	Set of activities
O	Set of occurrences
T	Set of periods
C	Set of child
C_p	Choice's value for the preference p
I	Set of inactive preferences
L	Set of forbidden preferences
PC_p	Price per child for each preference
PF_p	Fixed price by activity
MA_t	Maximum number of assignations during period t
S	Set of similarity labels

self explanatory. The cost for the organizers is composed of a fixed cost that remains the same regardless of the number of children and a price per child. Set PC_p represents the fixed cost, whereas PF_p represents the variable cost. Each activity is part of a group of similar activities with the same reference. Similar activities are given a same label. This set of labels is noted S .

We modelled this problem as an integer problem, in which the objective function and the constraints are linearly defined.

A. Variables

A solution to the considered problem is a set of pairwise associations between a child and an occurrence. We considered each of these pairwise associations to get either a true value if the child is assigned to the occurrence, and false if not. We therefore defined a set of variables that describes valid assignment of an occurrence and a child:

$$x_i = \begin{cases} 1 & \text{if preference } i \text{ is selected} \\ 0 & \text{otherwise} \end{cases}$$

where $i \in P$.

A variable is associated to each element of the table *preference*. Therefore, x_i true means that the associated occurrence of an activity is assigned to the associated child. In other words, this is equivalent to defining the following two-indexes variables, but reduces the number of variables, as a preference exists only if a child chose an occurrence.

$$x_{ij} = \begin{cases} 1 & \text{if child } i \text{ gets occurrence } j \\ 0 & \text{otherwise} \end{cases}$$

where $i \in C, j \in O$. For example, for a child choosing 4 occurrences between 10 possible occurrences, it will need 10 variables for the two-indexes variables, while only 4 variables are needed for the single-indexed model.

We also defined auxiliary variables that indicate if an occurrence is open or not. This was necessary since some occurrences can only be open if there is a minimum number of children to do the activity.

$$y_j = \begin{cases} 1 & \text{if occurrence } j \text{ is open} \\ 0 & \text{otherwise} \end{cases}$$

where $j \in O$.

B. Constraints

We considered all constraints as hard constraints.

- 1) Forbid assignment to cancelled occurrences.

This constraint is formalised with the assignment to a null value for each preference set to a cancelled occurrence. For efficiency reasons, it is advised to sum all such cases to zero.

$$\sum_{p \in I} x_p = 0 \quad (1)$$

- 2) Do not assign two incompatible occurrences to the same child.

$$x_a + x_b \leq 1 \quad \forall (a, b) \in \text{Incompatible} \quad (2)$$

- 3) Maximal number of activities per child and per period. This constraint avoids assigning too many activities during the same period to a child, even if their occurrences do not overlap.

Let's define, for each child c and each period t , the subset $CP_{c,t}$ of preferences for which the occurrences begin during period t . This subset $CP_{c,t}$ contains the preferences expressed by child c for activities occurring during period t .

$$\sum_{p \in CP_{ct}} x_p \leq MA_t \quad (3)$$

$$\forall (c, t) : c \in C, t \in T$$

- 4) At most 1 activity from a set of similar activities.

Let's define for each child c and each similarity value s the subset $SP_{c,s}$ of preferences with the same similarity label s for child c .

$$\sum_{p \in SP_{c,s}} x_p \leq 1 \quad (4)$$

$$\forall (c, s) : c \in C, s \in S$$

- 5) Forbid reassignment to previously assigned 'lifetime' activities.

The constraint consists in assigning a null value to each such preference.

$$\sum_{p \in L} x_p = 0 \quad (5)$$

- 6) Minimal age to perform an activity.

Let's define the set $BA \subset P$ of preferences indexes for which the child's age is below the minimum required age. We therefore forbid such assignments.

$$\sum_{p \in BA} x_p = 0 \quad (6)$$

- 7) Maximal age to perform an activity.

Let's define the set $AA \subset P$ of preferences indexes for

which the child's age is above the maximum age. We therefore forbid such assignments.

$$\sum_{p \in AA} x_p = 0 \quad (7)$$

- 8) Knome: group of children performing the same set of activities.

This constraint is split into two parts. In the first part, if members of a knome have chosen different occurrences, then each of them can not be accepted and therefore corresponding assignments get a value of zero. Let's define $D_p \in \{0, 1\}$ with value 1 if the corresponding occurrence is not chosen by the other member(s) of the knome, and value 0 otherwise.

$$\sum_{p \in P} D_p \cdot x_p = 0 \quad (8)$$

In the second part, the chosen occurrences should be either both accepted or both rejected; therefore, the corresponding variables must be equal.

$$x_i = x_j \quad (9)$$

$$\forall (i, j) \in \text{Knome}$$

- 9) Maximum number of children per occurrence.

Let's define for each occurrence o the subset PO_o of preferences that applies on occurrence o , and M_o the constant value 'maxchild' that indicates the maximal size of the assigned children's group. The following constraints specify the maximum size of the occurrence, and forbid assignment if the occurrence is not open.

$$\sum_{p \in PO_p} x_p \leq M_o \cdot y_o \quad (10)$$

$$\forall o \in O$$

- 10) Minimum number of children per occurrence to open it.

Let's consider again for each occurrence o the subset PO_o of preferences that applies on occurrence o . Define m_o as the constant value 'minchild' that indicates the minimal size of the assigned children's group. The constraint specifies that either the occurrence is performed by a minimum number of children, or it is not open.

$$\sum_{p \in PO_p} x_p \geq m_o \cdot y_o \quad (11)$$

$$\forall o \in O$$

- 11) Multi-occurrences activities.

This constraint is split into two parts. In the first part, an assignment gets a value of zero if it belongs to a multi-occurrences activity for which not all necessary occurrences have been chosen. In other words, incomplete activities can not be assigned to children. Let's define a

new set IV , with value 1 if such a case happens, and value 0 otherwise.

$$\sum_{p \in P} IV_p \cdot x_p = 0 \quad (12)$$

In the second part, all occurrences belonging to the same multi-occurrence activity should be either all accepted or all rejected; therefore, the corresponding variables have to be equal. Let's define PN as the set of corresponding preferences of successive occurrences.

$$x_p = x_{PN_p} \quad (13)$$

$$\forall p \in P$$

C. Objective function:

Several models exist to achieve the goals: maximizing the choices of the kids and minimizing the cost differences between the children. This could be solved via goal programming, via looking for a pareto optimal solution, or via converting the problem into a mono-objective one. In order to maximize the preferences of each child while minimizing the gap between the cost of each child, the function to maximize was defined as a pondered sum of the total deviation from the mean cost per child.

1) Preferences maximisation

Let nb_c be the number of choices that each child can express for each period. This means that the preferred occurrence gets the choice value 1, and the least preferred occurrence within the period gets a choice value of at most nb_c (a child might wish to select fewer occurrences per period than nb_c).

To mitigate disparities between preferences, we applied a power function to the preference weights and hence maximized the following z_{pref} function:

$$z_{pref} = \sum_{p \in P} x_p \cdot 2^{(nb_c - C_p)} \quad (14)$$

Without this power function, two sets of choices $\{1,1,1,4\}$ and $\{2,2,2,1\}$ would have the same assignation score of 9, although with this power function, the first one would receive a score of 25 and the second one a score of 20, favouring the assignation to the first choices.

2) Cost balancing

a) Cost minimization

Each activity has its own cost paid by the organizers. Children, however, pay the same price for the whole duration of Passeport Vacances, whether they get a helicopter flight or a museum visit. This means that the cost paid by the children does not depend on the assignment to activities and as there may be significant cost differences between

activities. The organizers would like to balance fairly the true cost between the children.

$$z_{price1} = \sum_{p \in P} x_p \cdot PC_p + \sum_{p \in O_p} y_o \cdot PF_p \quad (15)$$

b) Deviation cost minimization

The first fairness proposition minimizes the total cost of the organisation. The goal is now to minimize the deviation from the mean cost per child as follows:

$$z_{price2} = \sum_{c \in C} \left| \frac{z_{price1}}{|C|} \cdot cst - cost_c \right| \quad (16)$$

Where N_{child} is the total number of children, cst is a constant fixed to 1.2 to allow a small margin to the average, and $cost_c$, representing the total assignment cost of each child, id defined as follows:

$$cost_c = \sum_{p \in PR_c} x_p \cdot PC_p + \frac{x_p \cdot PF_p}{\sum_{p_1 \in O_p} x_{p_1}} \quad (17)$$

$\forall c \in C$

Where PR_c is the set of all the preferences of child c .

The second part of this sum is not linear. We approximated this sum by dividing by the maximum allowed number of children for the preference (MC_p) as follows:

$$cost_c = \sum_{p \in PR_c} x_p \cdot PC_p + \frac{x_p \cdot PF_p}{MC_p} \quad (18)$$

The linearization of the absolute value is a well-known technique (see for example Ünal et Uysal [2]). Adding two real variables, c^- representing the lack and c^+ representing the surplus, we can rewrite as follows:

$$\frac{z_{price}}{|C|} \cdot cst - cost_c = c_c^+ - c_c^- \quad (19)$$

$$\text{for } c \in C$$

And the following constraints:

$$c_j^+ \geq 0 \text{ for } j \in C \quad (20)$$

$$c_j^- \geq 0 \text{ for } j \in C \quad (21)$$

Finally, we chose to minimize only the cost above the average cost. The idea here is not to privilege a child with low cost assignment ($c^-[c]$), but to penalize the costs more highly ($c^+[c]$).

$$z_{price} = \sum_{c \in C} c_c^+ \quad (22)$$

Therefore, the objective function is defined as

$$z = w_{pref} \cdot z_{pref} - w_{price} \cdot z_{price} \quad (23)$$

TABLE X
SUMMARY OF THE OBJECTIVES FUNCTIONS

	w_{pref}	w_{pice}	z_{price}	z
Preference maximization	1	0	-	(14)
Cost minimization	2	1	(15)	(23)
Deviation cost minimization	2	1	(22)	(23)

TABLE XI
AMOUNT OF DATA USED TO TEST THE METHODS

634 Child	16621 preferences
1121 activity	50 knomes
532 occurrences	3 periods

The following weights have been defined in the purpose of fixing the importance of each objective by normalizing each one between 0 and 1 and multiplying by a factor to weight each objective relatively.

$$w_{pref} = 2 / \sum_{p \in P} x_p \cdot 2^{(nbc - C_p)} \quad (24)$$

$$w_{cost} = 1 / \sum_{p \in P} PC_p + \sum_{o \in O} PF_o \quad (25)$$

We chose to fix the importance of the preference maximization as 2 times that of the cost balancing.

Table X summarizes our three distinct objective functions: preference maximization, which focuses only on the maximization of the preferences; cost minimization that maximizes the preferences and minimizes the costs; and deviation cost minimization that maximizes the preferences and minimizes the deviation from the mean cost.

IV. RESOLUTION

The implementation of our exact LP model was done via the open source Julia 0.6.2 language [8], [9] and the MIP solver Gurobi 0.3.3 [10]. Tests were carried out on a 3.2 GHz Intel Core i5 CPU computer with 4 GB RAM, running 64-bit Ubuntu 16.04 LTS. Julia is defined as a "high-level, high-performance dynamic programming language for numerical computing" [11]. It allows, among other things, distributed parallel execution and shows a very good performance compared to the C language. Gurobi is a state-of-the-art MIP solver. The amount of data for this real-world problem is presented in Table XI. Compared to the currently used 20-year old software, based on an iterative heuristic, time to solve the problem was drastically reduced from 11 hours to less than 5 minutes on similar single core computers.

V. EVALUATION

To evaluate the balancing, both methods, minimization of the total cost and minimization of the deviation from the mean cost, were tested with real data and compared to the solution obtained from an objective function without any fairness component.

Unsurprisingly, the cost minimization method obtains the smallest mean price, but does not improve the value of the

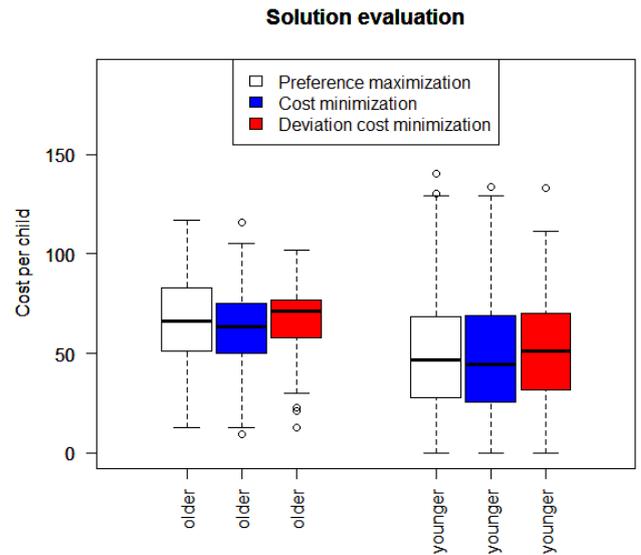


Fig. 1. Comparison of cost's repartition between children

standard deviation or the coefficient of variation compared to the preference maximization. Moreover, cost minimization decreases the total number of assignments, which goes against the desired goal. The second balancing form, deviation cost minimization, has a more equitable repartition of the costs between children. The standard deviation is reduced by 18% for the group of older children and by 25% for the younger group, which has the smallest coefficient of variation. The number of assignments is slightly higher, at the price of a small decrease in children's preferences. We also note that adding new variables for the approximation of the absolute value penalizes the performance, as shown in Table XIV, but it still remains reasonable in this context. Table XII and Figure 1 present a summary of the obtained solutions. Column cv corresponds to the coefficient of variation and column threshold refers to outliers detection (i.e. a threshold value beyond which a cost is considered as an outlier). Finally, Table XIII presents the number of assignments by order of choice's priority: in a fair solution, fewer first choices are assigned to children, replaced by second or third choices.

VI. CONCLUSION

This article presents the modelization of a real-world assignment problem in which the objective is to assign children to activities they chose respecting as much as possible their preference order without violating the different constraints. The second part of the study focuses on the desire to balance the costs fairly between children, as they all pay the same subscription cost to Passeport Vacances. Two methods are proposed, which take into account the costs and are compared with the solution without cost balancing. The cost's minimization deteriorates the quality of the assignment without decreasing the inequality between each children, while the

TABLE XII
SUMMARY OF THE ASSIGNATION SOLUTION FOR EACH METHOD

	Age category	Mean cost	Std cost	Nb child	Threshold	cv
Preference maximization	1	49.60	29.37	570	131.83	0.59
	2	65.32	25.40	51	136.45	0.39
Cost minimization	1	48.31	29.39	570	130.60	0.61
	2	62.43	25.93	51	135.05	0.42
Deviation cost minimization	1	49.83	24.17	570	117.51	0.49
	2	65.31	18.87	51	118.16	0.29

TABLE XIII
NUMBER OF ASSIGNATION, ORDERED BY CHOICE'S PRIORITY FOR EACH OF THE 3 METHODS

Choices	Nbr of assignments	Nbr of assignment	Nrb of assignments
	Preference maximization	Cost minimization	Deviation cost minimization
1	3372	3375	3369
2	1014	1011	1028
3	414	392	416
4	173	156	172
total	4973	4934	4985

TABLE XIV
TIME OF RESOLUTION FOR EACH MODELS

Time of assignment	Time of assignment	Time of assignments
Preference maximization	Cost minimization	Deviation cost minimization
2 minutes and 7 seconds	2 minutes and 7 seconds	3 minutes and 27 seconds

minimization of the deviation from the mean price gets closer to the desired goal, but increases the computational time because of the increase in the number of variables. Improving cost balancing without exploding the number of variables or computational time, by means of valid inequalities, is one possible future research direction.

REFERENCES

[1] F. Uney-Yuksektepe and İ. Karabulut, "Mathematical programming approach to course-teaching assistant assignment problem," in *Proceedings of the 41st International Conference on Computers & Industrial Engineering*, 2011, pp. 878–883.
 [2] Y. Z. Ünal and Ö. Uysal, "A new mixed integer programming model for curriculum balancing: Application to a turkish university," *European Journal of Operational Research*, vol. 238, no. 1, pp. 339–347, 2014. doi: <https://doi.org/10.1016/j.ejor.2014.03.015>
 [3] R. de la Torre, A. Lusa, and M. Mateo, "A MILP model for the long term academic staff size and composition planning in public universities," *Omega*, vol. 63, pp. 1–11, sep 2016. doi: [10.1016/j.omega.2015.09.008](https://doi.org/10.1016/j.omega.2015.09.008)
 [4] E. L. Bouzarth, R. Forrester, K. R. Hutson, and L. Reddoch, "Assigning students to schools to minimize both transportation costs and socioeco-

conomic variation between schools," *Socio-Economic Planning Sciences*, sep 2017. doi: [10.1016/j.seps.2017.09.001](https://doi.org/10.1016/j.seps.2017.09.001)
 [5] B. Domenech and A. Lusa, "A MILP model for the teacher assignment problem considering teachers' preferences," *European Journal of Operational Research*, vol. 249, no. 3, pp. 1153–1160, mar 2016. doi: [10.1016/j.ejor.2015.08.057](https://doi.org/10.1016/j.ejor.2015.08.057)
 [6] L. Pan, S. Chu, G. Han, and J. Z. Huang, "Multi-criteria student project allocation: A case study of goal programming formulation with dss implementation," in *The Eighth International Symposium on Operations Research and Its Applications (ISORA'09), Zhangjiajie, China, 2009*, pp. 75–82.
 [7] I. Kim and O. de Weck, "Adaptive weighted-sum method for bi-objective optimization: Pareto front generation," *Structural and Multidisciplinary Optimization*, vol. 29, no. 2, pp. 149–158, sep 2004. doi: [10.1007/s00158-004-0465-1](https://doi.org/10.1007/s00158-004-0465-1)
 [8] J. Bezanson, S. Karpinski, V. B. Shah, and A. Edelman, "Julia: A fast dynamic language for technical computing," *CoRR*, vol. abs/1209.5145, 2012. [Online]. Available: <http://arxiv.org/abs/1209.5145>
 [9] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017. doi: [10.1137/141000671](https://doi.org/10.1137/141000671). [Online]. Available: <http://julialang.org/publications/julia-fresh-approach-BEKS.pdf>
 [10] [Online]. Available: <http://www.gurobi.com/>
 [11] [Online]. Available: <https://julialang.org>

Optimizing Maintenance in Project Management by Considering Health, Safety, Environment and Resilience Engineering

Behnam Einabadi
School of Industrial Engineering
University of Tehran
Tehran, Iran
behnam.einabadi@ut.ac.ir

Pedram Memari
School of Industrial Engineering
University of Tehran
Tehran, Iran
memari.pedram@ut.ac.ir

Seyed Farid Ghaderi
School of Industrial Engineering
University of Tehran
Tehran, Iran
ghaderi@ut.ac.ir

Abstract— One of the most important objectives of project management is to complete the project within the specified completion date of the project. Another important objective of project is to terminate the project by minimum rate of injuries and damage to the environment. One of the important factors which affect the time objective of the project is the failures or breakdowns of the project Machines and Equipment. Also, Health, Safety, Environment (HSE) factors are crucial in the efficient execution of the project. Resilience engineering is a new concept that will improve the safety and reliability of a high-risk system such as power plant construction project. Previous studies didn't consider the resilience engineering (RE) factors which could help the project to achieve its goals. Related data was collected from a power plant construction project and fuzzy DEA and Z-number DEA were utilized to analyze the Data and Best DEA model is selected according to maximum average efficiency and also for identifying most effective factors sensitivity analysis was done, and we found that flexibility, and project percent progress, system downtime, reporting culture and HSE costs are the most important factors on maintenance of the project. To the best of our knowledge, this is the first study considering RE and HSE factors to optimize maintenance of the project.

I. INTRODUCTION

RECENTLY the importance of RE has been pointed by numerous studies in order to analyze effective factors of RE in the system to able to recover to its initial/proper state. In this study construction project of a power plant has been considered as a system and beside the RE factors, related HSE factors which is crucial for certifying health, safety and environment issues in the construction site and their effects on the progress/maintenance of project have been considered and analyzed in order to find the most effective factors that improving them, could result in a project with increased level of resilience and also better progression of the project.

One of the goals of the construction project is to be finished as soon as the planned date, since the role of humans and personnel is crucial in achieving the objectives, especially who are dealing with executive operation in the construction site, who are working in an environment with plenty of risks, safety management tool became one good method to control

and limit the incidents. It could minimizing accidents. Safety culture could help the project to reach to its targets, also resilience engineering (RE) is a novel approach that could control and limit incidents and accidents in the high-risk environment, in the previous studies RE is not considered. DEA models were employed to analyze the RE factors in a power plant construction project. One of the contribution of this paper is the incorporation of RE factors and DEA analysis to identify prominent effecting project factors and take new strategies in the future construction projects of the same type.

This study analyzes the defined HSE and RE factors of a construction project of a combined cycle power plant on the maintenance of the case project by DEA tools so that by the optimization models of fuzzy DEA and ZDEA, the most effecting factor and the most productive month of the project is defined, for taking new strategies in the next projects in a way that failures and accidents become low and lower and the maintenance of the project become high and higher.

This study is in a phase of construction of such a plants that has its own issues, problems and difficulties, as a case study we analyzed our research in a combined cycle power plant in Yazd-Iran. Also in this study we consider HSE factors, Resilience engineering factors altogether and we assess them on the maintenance of the project.

Definition of the RE, HSE and maintenance factors has been pointed in the following of introduction, the rest of this paper has been constructed as following: methodology and the structure of our study in section II, explanation of the case study and data analysis in section III, results and discussion and sensitivity analysis in section IV and conclusion in section V.

A. Resilience Engineering

Three system states in the operation of an industrial process can be distinguished as catastrophic, upset and normal ones. Project-oriented companies try to keep system in normal state and to achieve this aim through the manipulation of operation variables. Whenever accidents or incidents happens. RE can help the system to recover from catastrophic or upset state to normal state[1].

Resilience is defined as “the ability of an organization (system) to keep or recover quickly to a stable state, allowing

it to continue operations during and after a major mishap or in the presence of continuous significant stresses” [2].

Here is some prominent studies described the concept and features of systems which are resilient [3], [4]. In order to increase the resilient level of the systems four factors has developed by Azadeh and Salehi [5]. In current study, six principles proposed by Hollnagel and Woods [6] and three principles which are Self-Organization, Team work and Redundancy and introduced by Azadeh et al. [7] will be used. These nine principles described as follows:

- **Management commitment:** Senior management perceps problems and hardships of personnel especially those related to safety issues, and attempts to solve them [2].

- **Reporting culture:** Identify the whole context and atmosphere of the project in which personnel feel free to report safety issues [8].

- **Learning:** Getting lessons from the normal state of work and abnormal events like accidents, is an important point of view in RE and there is quite emphasis on that [8].

- **Awareness:** management will be aware of what happens in the construction site by data that are collected [9].

- **Preparedness:** Problems and hardships that are in result of safety issues, human performance and equipment breakdowns is predicted by the project team, and they becomes ready in order to response [9].

- **Flexibility:** When an unexpected event occurs, if the organization is agile enough to response the event, whether by using on hand resources or other external resources, this organization called to be flexible [9].

- **Self-organization:** Self-organization happens when the authority is distributed in project personnel [10]. In such systems, there are interdependent entities issuing orders which collaborate with each other and share information and try to adjust themselves to the feedback of other agents [11]. These kind of systems normally conquer a wide range of faults and variation [10].

- **Teamwork:** productivity, adaptability and job satisfaction will increase when teamwork becomes to work [12]. Due to the individual and organizational pressures RE basis says that human errors are unavoidable [13]. Teamwork is on the basis of mutual support, communications, leadership, and situation monitoring [14]. When there are a pile of tasks, if the staff assist and support each other, organizational, individual pressures and human errors will be reduced, and as a result reliability and safety will be improved [7].

- **Redundancy:** Redundancy means the availability of alternative pathways of resources such as Equipment, machines and manpower in order to respond and use when the basic parts and elements such as project equipment is inaccessible and unavailable [15]. In order to develop such systems, required elements in the case of disruptions, disturbances, and non-normal conditions, must be procured and be on hand in advance [16]. Redundancy in current study, has been considered in resources like equipment, machinery and manpower.

B. Health, Safety and Environment

Due to sensitiveness of Health Safety and Environment (HSE) issue, they have to be considered priori than any other subject and nothing can equal to be important as protection of human health, safety and the environment [17]. It is unfortunately truth that the forgotten right of all the workers and employees is to work safely in an environmentally responsible manner.

One of the most important objectives of a construction project must be terminating the project with minimum injuries and minimum damage to the environment. In this study seven factor has been considered as HSE indexes, these are: Checkup and Examination, Issue Health Card for Personnel, Instruction, Identified dangers and assessment of the risks, HSE Encouragement, HSE Caveat, HSE Costs and gasoline usage.

Checkup& examination and issue health card for personnel are related to health factor that are the data of periodic health checkup of the project personnel and the health card indicates the data of the sensitive personnel which they may affect the health of other personnel.

For safety factor following indexes has been considered:

Instruction: related data of the number of personnel, instructed by HSE experts before and during execution work.

Identified dangers and assessment of the risks: the number of HSE risks which has been identified by HSE experts and further assessed.

HSE Encouragement, HSE Caveat: are the recorded number of encouraging personnel for their commitment in practice and on the other hand vice versa.

HSE cost: related HSE costs has been recorded.

Gasoline usage: the amount of used gasoline for execution of the project like usage of the cranes, are recorded.

C. Maintenance

“The main purpose of maintenance is to retain systems in or to restore them to a functioning state. Maintenance also contributes to improved system knowledge and inter-discipline coordination that may benefit the entire organization” [18]. The growing complexity and significance of the projects and the importance of completing the project within its planned schedule has made us to consider the maintenance program in project execution. One of the most important of the reason is the availability and reliability of project equipment, machinery and manpower is vital for keep the project in the functioning state [19].

In this study the factor of project system down time which means the time which the project has been hold due to equipment issues or accidents or incidents due to HSE affairs or any unpredicted element, affected the maintenance of the project, and the project progress percent in different month of the project are considered as a maintenance factors.

Most of studies are performed in a plant that is currently in process and maintained [7, 21, 22], while this study is performed in a construction phase of such plants that has its own issues, problems and difficulties. Also in this study we

consider HSE factors, Resilience engineering factors altogether and we assess them on the maintenance of the project. Features of this study with respect to previous studies is shown in TABLE I. Methodology

In this study, effective maintenance factors are identified in a combined cycle power plant construction project by HSE and RE and Z-number DEA. The most crucial factors which will affect the maintenance of the project will be available by considering and modeling of HSE and RE factors according to the process which is depicted in Fig 1.

In most industrial projects in Iran, there is a system of project management, and within this kind of system the data relating

to the project and HSE affairs are being recorded by headquarters of the project team. There are different sub-contractors that are managed by a general contractor. According to defined factors for this study in previous section HSE and maintenance factors are considered by the available data in different month of the project execution life cycle. Different month of the project which is 38 month in this study are considered to be the decision making units. Relating resilience data are considered by interviewing a group of managerial board. In order to, certify reliability of the data, alpha Cronbach is computed by SPSS®.

TABLE I.
FEATURES OF THIS STUDY VERSUS OTHER STUDIES AND METHODS

	Resilience factors	HSE factors	Maintenance factors	Project oriented company	Practicality in real world	Statistical method	Identification of important factors	Sensitivity analysis
Azadeh et al. [20]	✓		✓		✓	✓		
Azadeh et al. [21]	✓				✓	✓	✓	✓
Shirali et al. [22]	✓	✓			✓	✓		
Azadeh et al. [7]	✓	✓			✓	✓	✓	✓
This study	✓	✓	✓	✓	✓	✓	✓	✓

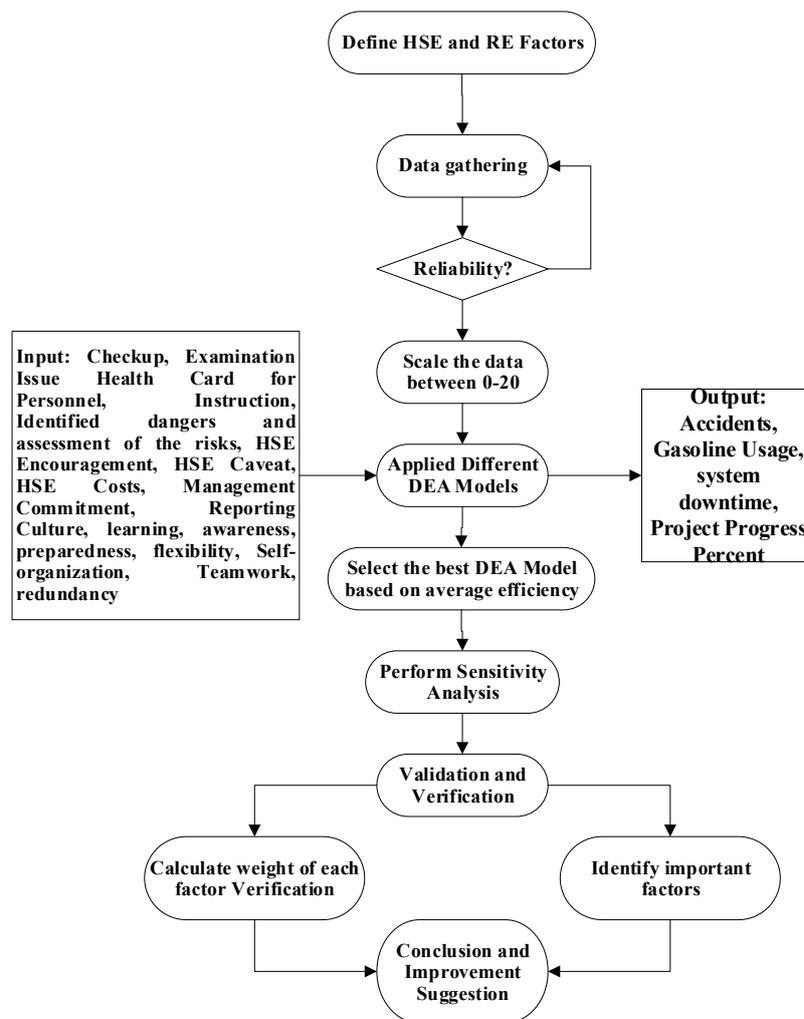


Fig 1. Schematic View of the Proposed Approach

D. FDEA

In real world problems, when the input and output values of the data are vague and not exact, some fuzzy models have been presented for dealing with uncertainty in the data envelopment analysis. Fuzzy DEA is one of the methods that output and input variables are asymmetrical triangular shaped fuzzy numbers and there is a lower bound and upper bound for the input and output variables of DMU values[23]. Consider $\tilde{x}_{ij} = (x_{ij}^p, x_{ij}^m, x_{ij}^o)$ fuzzy values of the input variable and $y_{ij} = (y_{ij}^p, y_{ij}^m, y_{ij}^o)$ fuzzy values of the output variable which have three pessimistic, mean and optimistic values. In this model, α cuts from the interval $[0, 1]$ are the parts of fuzzy sets that generates regular sets, and by each α , linear programming model is used in order to optimize solution[24]. The model is on the following Eq (1). In our problem there are 38 number of DMU's, 16 number of input factors and 4 number of output factor. The values for pessimistic and optimistic of the collected data are considered as following:

Minimum value of the each factor from its data is divided by two and the value is subtracted from the corresponding value of the DMU for Pessimistic state and it is added for optimistic state.

Optimum value of θ refers to efficiency of the DEA model.

Indices

i	Indices of DMUs
j	Indices of inputs
r	Indices of outputs
n	Number of DMUs
m	Number of inputs
s	Number of outputs
DMU(i)	The i th DMU
DMU(0)	The target DMU ($i = 0$)

Parameters

\tilde{Zx}_{ji}	Z-number value of input j related to DMU i
\tilde{Ax}_{ji}	Fuzzy value of input j related to DMU i
\tilde{Bx}_{ji}	Fuzzy reliability value of input j related to DMU i
\tilde{Zy}_{ji}	Z-number value of output r related to DMU i

Variables

λ_i	Weight variables in the proposed model for obtaining the efficiencies of DMUs
θ_0	Objective value (efficiency) of the DEA model

$$\text{Min } \theta \quad \text{Eq (1)}$$

s.t.

$$\theta(\alpha x_{ip}^m + (1 - \alpha)x_{ip}^p) \geq \sum_{j=1}^{38} \tau_j (\alpha x_{ij}^m + (1 - \alpha)x_{ij}^p) \quad i = 1, \dots, 16$$

$$(\alpha y_{rp}^m + (1 - \alpha)y_{rp}^p) \leq \sum_{j=1}^{38} \tau_j (\alpha y_{rj}^m + (1 - \alpha)y_{rj}^p) \quad r = 1, \dots, 4$$

$$\sum_{j=1}^{38} \tau_j = 1,$$

$$\tau_j \geq 0 \quad j = 1, \dots, 38$$

E. Z-number DEA

The concept of Z-numbers DEA was introduced by Zadeh [25] which is dealing with reliability of information and it consist of two parts $Z = (A, B)$. A is a fuzzy number of the variable which is described in previous part and B is the extent of the reliability of each of three A values. The extent of reliability could be the amount of sureness, believes of people about a phenomenon, etc. a theorem have been proven by Kang et al [26] which transforms Z-number to normal fuzzy set. In this paper the CCR model and the MATLAB coded for proposed model that is developed by Azadeh and Kokabi [27] is used to optimize solution. In this model $\tilde{Zx}_{ji} = (\tilde{Ax}_{ji}, \tilde{Bx}_{ji})$, \tilde{Ax}_{ji} is the triangular fuzzy number and \tilde{Bx}_{ji} is the certainty measure of \tilde{Ax}_{ji} . The structure of the CCR model are presented in Eqs (2) and (3).

$$\text{Min } \theta_0 \quad \text{Eq (2)}$$

s.t.

$$\sum_{i=1}^n \lambda_i \tilde{Zx}_{ji} \geq \theta_0 \tilde{Zx}_{j0} \quad j = 1, \dots, m$$

$$\sum_{i=1}^n \lambda_i \tilde{Zy}_{ri} \geq \tilde{Zy}_{r0} \quad r = 1, \dots, s$$

$$\lambda_i \geq 0 \quad i = 1, \dots, n$$

$$\text{Max } \theta_0 = \sum_{r=1}^s u_r \tilde{Zy}_{r0} \quad \text{Eq (3)}$$

s.t.

$$\sum_{j=1}^m v_j \tilde{Zx}_{j0} = 1$$

$$\sum_{r=1}^s u_r \tilde{Zy}_{ri} - \sum_{j=1}^m v_j \tilde{Zx}_{ji} \leq 0, \quad i = 1, \dots, n$$

$$u_r, v_j \geq 0 \quad r = 1, \dots, s, \quad j = 1, \dots, m$$

II. CASE STUDY

The study has been implemented on a combined cycle power plant project in Yazd-Iran. The cooling system of the power plant is kind of ACC which is suitable for dry climates. The combined cycle part has been synchronized on November 2016. About 1,337,830 Man-hour had been working throughout the project. An integrated strict system of HSE had been implemented through the construction site by MAPNA Company which in this system the statistics of the project was being recorded, and during the life cycle of the project there was no incident which lead to death. According to the previous section required data collected from MAPNA Company.

In our case study, models of DEA are utilized in order to recognize the performance of RE and HSE in a power plant project. Factors such as Checkup, Examination Issue Health Card for Personnel, Instruction, Identified dangers and assessment of the risks, HSE Encouragement, HSE Caveat, HSE Costs, redundancy, Management Commitment, preparedness, awareness, flexibility, learning, Self-organization, Reporting Culture, Teamwork, are considered as Input and four factor which are Accidents, Gasoline Usage,

system downtime, Project Progress Percent are considered to be output of the model.

The important objective of current study is to evaluate maintenance of different month of the project and to determine the most important factors in overall performance.

III. RESULTS AND DISCUSSION

Fuzzy DEA and Z-number DEA are effective methodology for analyzing the efficiency of project performance in different month which are considered as DMUs. Related data to the issues of safety, health and environment and project maintenance has been collected from a combined cycle power plant, according to the experience of the writer and interviewing from project staff, nine factor of RE have been weighted in different month of the case project, due to the uncertainty of the weights and possible mistakes in recording data, we analyzed via FDEA and Z-number DEA in order to find the best month as an effect pattern and also determine the important factors that investment in such factors in the future projects could increase the resilient and maintenance level of the same type of projects.

A. Reliability Test on Data

The data achieved, are analyzed in SPSS® software. Using Cronbach’s alpha via SPSS®, the Data has been analyzed and Cronbach’s alpha’s value is 87% that is quite acceptable. TABLE II is the output of the SPSS® software.

TABLE II. SPSS RESULT

Cronbach's Alpha	N of Items (Number of data)
0.869	20

B. Result of Fuzzy DEA and Z-number DEA

Collected Data is weighted for performing Fuzzy DEA and Z-number DEA, the following weights in TABLE III has been used for applying Z-number DEA. All the data are scaled into [0, 20] since we could see better the difference between results and the whole result and data are according to this scale. By applying the methods, efficiency scores of both methods for different alphas are computed, following result is shown in Fig 2.

Optimum alpha for both models are considered base on maximum average efficiency, for Fuzzy DEA optimum alpha is 0.05 and for Z-number DEA is 0.1, TABLE IV shows the result of the models, and also TABLE V shows the corresponding average efficiency values.

TABLE III. Z-NUMBER WEIGHTS

Sure	[15,17.5,20]
Usually	[10,12.5,15]
Likely	[0,5,10]

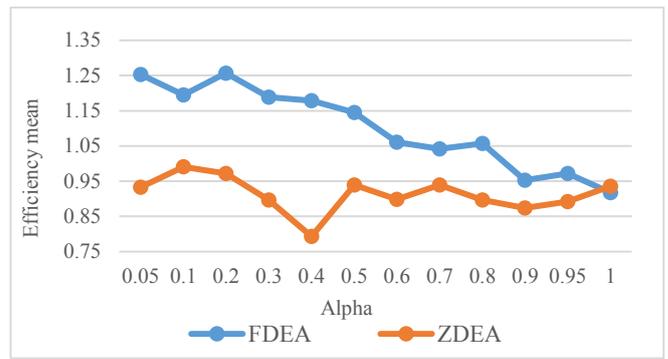


Fig 2. Alpha Results

TABLE IV. FDEA & Z-DEA RANKING

DMU (Month)	FDEA		Z-Number DEA	
	$\alpha=0.05$	Rank	$\alpha=0.01$	Rank
1	1.4955	20	1.2138	16
2	1.6947	4	1.3054	1
3	1.7321	2	1.2833	3
4	1.7643	1	1.3022	2
5	1.7171	3	0.0000	35
6	0.0000	35	1.2687	4
7	1.4872	21	1.1992	21
8	1.5736	10	1.2155	12
9	1.5543	15	1.2188	11
10	1.5736	10	1.2143	14
11	1.5044	19	1.1562	23
12	0.0000	35	1.2316	9
13	1.6165	7	0.0000	35
14	0.7924	32	0.6086	33
15	1.4428	23	1.2082	19
16	1.6007	9	1.2233	10
17	1.0727	29	0.9732	28
18	1.5736	10	1.2150	13
19	1.3863	24	0.0000	35
20	0.0000	35	0.0000	35
21	0.8878	31	0.7801	30
22	1.6230	5	1.2417	7
23	1.5515	16	1.1654	22
24	1.5343	18	1.1999	20
25	0.0000	35	1.2138	17
26	1.5420	17	1.2138	15
27	1.1261	27	0.9841	27
28	1.1024	28	0.9688	29
29	1.2980	25	1.0660	25
30	1.1923	26	1.0534	26
31	1.4472	22	1.1395	24
32	1.6021	8	1.2331	8
33	1.5667	14	1.2106	18
34	0.7470	33	0.6653	32
35	1.6230	5	1.2424	6
36	1.5729	13	1.2540	5
37	0.7045	34	0.5023	34
38	0.8957	30	0.6690	31

TABLE V. AVERAGE EFFICIENCY

FDEA	Z-Number DEA
1.25258	0.99054

For developing Fuzzy model, min and max of each DMU is considered to be half of minus and plus of the minimum of the every factor values. And following weights are considered for Z-number DEA:

TABLE IV shows the efficiency of each DMU. As previously stated, DMU’s are the different month of the

project, and the value of efficiency shows performance of each month that the higher the ranks are the more is efficient month in case of progress and maintenance.

C. Noise and Result

In order to reinforce the certify the result of the Fuzzy DEA, we have analyzed the Model with exertion of the noise in the data, which we randomly selected 30 data from the rows and columns of the data in the corresponding value has been changed with faraway values and again Fuzzy DEA model has been run for different α cuts, the results are shown in Fig 3.

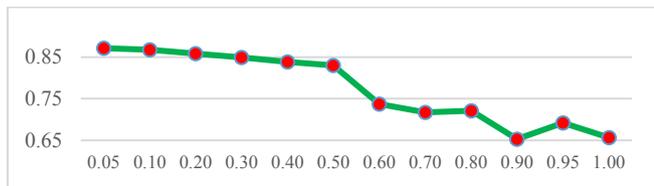


Fig 3. Noised FDEA

As it is shown in above figure, again we see that the maximum average efficiency occurs in the α cut of the 0.05.

The spearman test between the result of Fuzzy DEA and Noised Data fuzzy DEA has been run and the result for different α cuts are shown in Fig 4.



Fig 4. Spearman test

D. Sensitivity Analysis

For performing sensitivity analysis, the prepared Table of data with the upper and lower bound values, and the weights $\bar{B}x_{ji}$ each of the columns of the factors removed once. The fuzzy DEA and Z-number DEA model was run with the α that has a maximum average efficiency, in order to calculate the efficiency of the model in the absence of that factor, if the average efficiency become less, it shows that the omitted factor has significant effect on the entire model, and respectively vice versa. The following result are in Table VI. The most effective factors on FDEA analysis are HSE encouragement, system downtime, preparedness, awareness, HSE costs and HSE caveats. The most effective factors are shown in Fig 5. The most effective factors on Z-number DEA are flexibility, project progress, system downtime, reporting culture and HSE costs. The most effective factors are shown in Fig6.

TABLE VI. AVERAGE EFFICIENCY ANALYSIS

Factor (Eliminated)	FDEA	Z-number DEA y	Factor (Eliminated)	FDEA	Z-number DEA y
Checkup, Examination	1.33304	1.01696	Awareness	1.15376	1.02737
Issue Health Card for Personnel	1.28893	1.04400	Preparedness	1.14570	0.99797
Instruction	1.19265	0.97528	Flexibility	1.27583	1.08762
Identified Dangers and Assessment of the Risks	1.22779	0.92095	Self-Organization	1.29249	0.95619
HSE Encouragement	1.38442	1.03127	Teamwork	1.31924	1.02213
HSE Caveat	1.34780	0.98304	Redundancy	1.19714	0.92349
HSE Costs	1.35043	1.06603	Accidents	1.28589	1.00704
Management Commitment	1.24152	0.99154	Gasoline Usage	1.25429	0.96959
Reporting Culture	1.25406	0.91063	System Downtime	1.13223	0.90549
Learning	1.28028	0.99110	Project Progress	1.18154	0.89802
			Percent Per month		

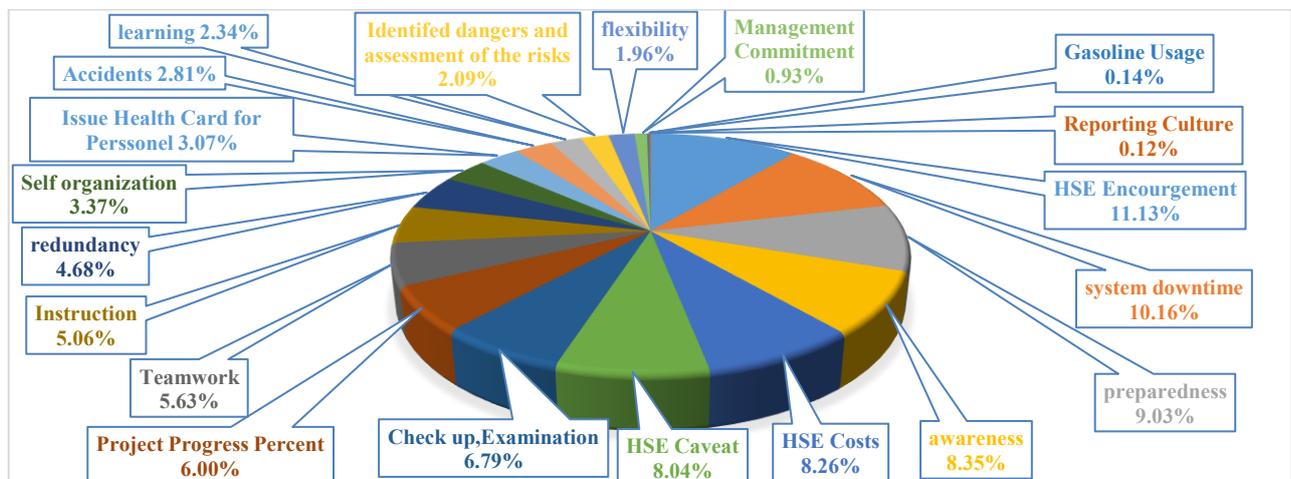


Fig 5. Weight of Factors in fuzzy DEA model

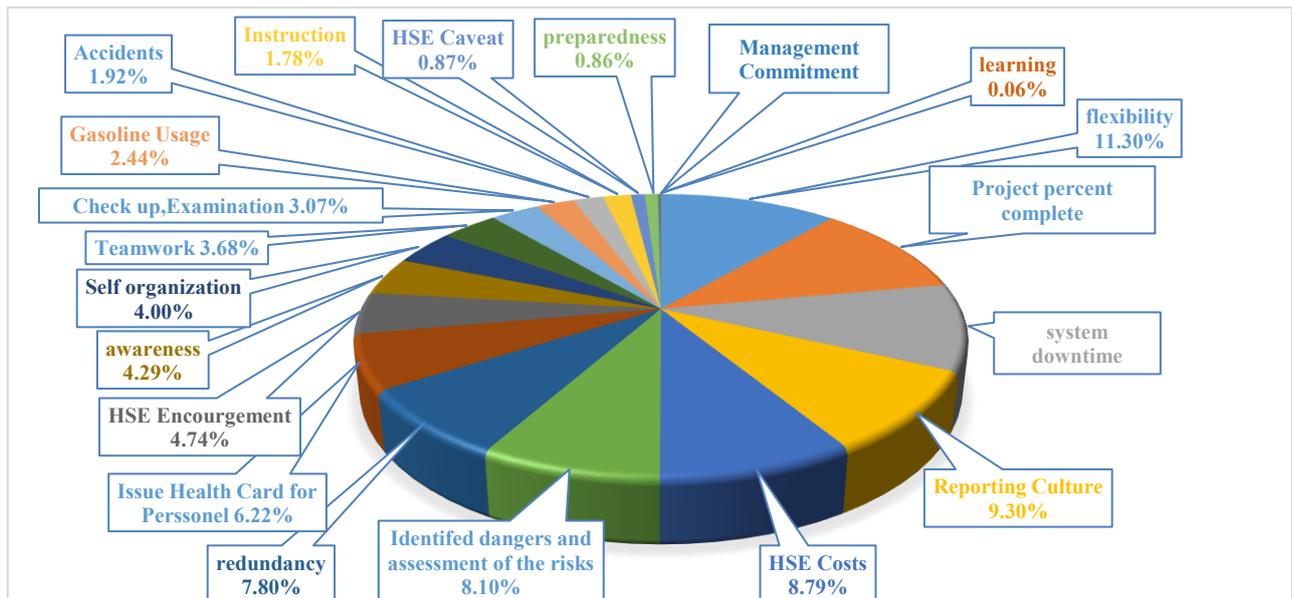


Fig 6. Weight of Factors in Z-number DEA

E. Validation and Verification

For validation of z-number model the correlation between the average efficiency of Fuzzy DEA and Z-number DEA are calculated by MINITAB®. The correlation is computed by spearman value between different alphas of the both models. The results are depicted in Fig 7.

Since the optimum alpha is 0.1 for Z-number method, correlation between Fuzzy DEA for this alpha is .721 so the value is substantial and the model is verified.

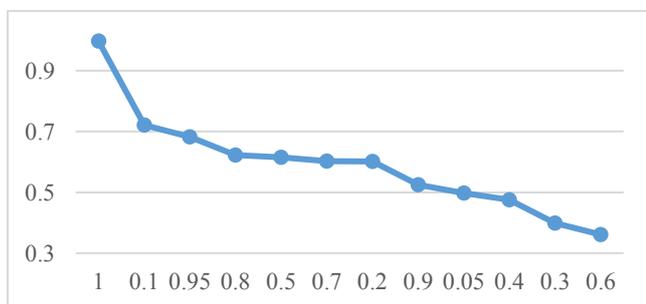


Fig 7. Correlation (Spearman value)

IV. CONCLUSION

Resilience engineering (RE) is a novel approach for safety improvement of highly risk systems such as Power plants and construction projects. This study identifies Resilience Engineering, HSE and maintenance factors of a power plant construction project by Z-number DEA. We believe to our knowledge, this is the first study examines the Resilience and HSE factors with respect to maintenance of the project by Z-number DEA. For do this, related data was collected from MAPNA Company in one of its project. Then DEA methods are applied to assess different factors effecting on maintenance of the project. According to higher average of efficiencies best DEA model selected. DMU efficiencies are

related to efficiency of the project in different months, which the sixteenth month has the highest rank that maintenance scores was better than other month of the project. Sensitivity analysis is used to determine the important factors, the results in this case of project show that flexibility, and project percent progress, system downtime, reporting culture and HSE costs are the most important factors on maintenance of the project. Giving more attention to improve flexibility, project progress, reporting culture and more investment in HSE cost and also decreasing system downtime would result in better maintaining the project and decrease the project time period. And we recommend that creating a Maintenance planning in the construction phase of the project could considerably mitigate the failures and increase the efficiency of the system.

REFERENCES

- [1] L. T. T. T. Dinh, H. Pasman, X. Gao, and M. S. Mannan, "Resilience engineering of industrial processes: Principles and contributing factors," *J. Loss Prev. Process Ind.*, vol. 25, no. 2, pp. 233–241, Mar. 2012.
- [2] J. Wreathall, "Properties of Resilient Organizations: An Initial View," in *Resilience engineering: Concepts and precepts*, Ashgate, Aldershot, UK, 2006, pp. 275–288.
- [3] N. G. L. Erik Hollnagel, David D. Woods, *Resilience Engineering: Concepts and Precepts*. Ashgate Publishing, Ltd., 2006.
- [4] C. P. Nemeth, E. Hollnagel, and S. W. A. Dekker, "Resilience Engineering Perspectives: v 2 Preparation and Restoration," p. XIX-288, 2009.
- [5] A. Azadeh and V. Salehi, "Modeling and optimizing efficiency gap between managers and operators in integrated resilient systems: The case of a petrochemical plant," *Process Saf. Environ. Prot.*, vol. 92, no. 6, pp. 766–778, Nov. 2014.
- [6] E. Hollnagel and D. D. Woods, "Epilogue: Resilience engineering precepts," *Resil. Eng. Precepts, ...*, no. January, pp. 347–358, 2006.
- [7] A. Azadeh, V. . Salehi, B. . Ashjari, and M. . Saberi, "Performance evaluation of integrated resilience engineering factors by data envelopment analysis: The case of a petrochemical plant," *Process Saf. Environ. Prot.*, vol. 92, no. 3, pp. 231–241, May 2014.
- [8] A. Azadeh, S. Motevali Haghghi, and V. Salehi, "Identification of

- managerial shaping factors in a petrochemical plant by resilience engineering and data envelopment analysis," *J. Loss Prev. Process Ind.*, vol. 36, pp. 158–166, Jul. 2015.
- [9] A. Azadeh and M. Sheikhalishahi, "An Efficient Taguchi Approach for the Performance Optimization of Health, Safety, Environment and Ergonomics in Generation Companies," *Saf. Health Work*, vol. 6, no. 2, pp. 77–84, Jun. 2015.
- [10] G. di Marzo Serugendo, "Robustness and dependability of self-organizing systems: a safety engineering perspective," in *Stabilization, Safety, and Security of Distributed Systems*, Proceedings, no. 5873, 2009, pp. 254–268.
- [11] D. A. Plowman, S. Solansky, T. E. Beck, L. Baker, M. Kulkarni, and D. V. Travis, "The role of leadership in emergent, self-organization," *Leadersh. Q.*, vol. 18, no. 4, pp. 341–356, Aug. 2007.
- [12] A. Xyrichis and E. Ream, "Teamwork: A concept analysis," *J. Adv. Nurs.*, vol. 61, no. 2, pp. 232–241, Jan. 2008.
- [13] J. Rasmussen, A. M. Pejtersen, and L. P. Goodstein, "Cognitive Systems Engineering John Wiley & Sons," Inc., New York, NY, USA, 1994.
- [14] J. Battles and H. King, "TeamSTEPPS® Teamwork Perceptions Questionnaire (T-TPQ) Manual," *Am. Inst. Res.*, pp. 23–25, 2010.
- [15] P. Kalungi and T. T. Tanyimboh, "Redundancy model for water distribution systems," *Reliab. Eng. Syst. Saf.*, vol. 82, no. 3, pp. 275–286, 2003.
- [16] F. Storseth, R. K. Tinmannsvik, and K. Øien, "Building Safety by resilient organization – a case specific approach," in *Reliability, risk and safety: theory and applications*, no. 1, 1997, pp. 1209–1214.
- [17] P. S. Gholami, P. Nassiri, R. Yarahmadi, A. Hamidi, and R. Mirkazemi, "Assessment of health safety and environment management System function in contracting companies of one of the petro-chemistry industries in Iran, a case study," *Saf. Sci.*, vol. 77, pp. 42–47, Aug. 2015.
- [18] P. Okoh and S. Haugen, "Improving the robustness and resilience properties of maintenance," *Process Saf. Environ. Prot.*, vol. 94, no. C, pp. 212–226, 2015.
- [19] B. Al-Najjar and I. Alsyof, "Improving effectiveness of manufacturing systems using total quality maintenance," *Integr. Manuf. Syst.*, vol. 11, no. 4, pp. 267–276, Jul. 2000.
- [20] A. Azadeh, M. S. Gharibdousti, M. Firoozi, M. Baseri, M. Alishahi, and V. Salehi, "Selection of optimum maintenance policy using an integrated multi-criteria Taguchi modeling approach by considering resilience engineering," *Int. J. Adv. Manuf. Technol.*, vol. 84, no. 5–8, pp. 1067–1079, Sep. 2016.
- [21] A. Azadeh, V. Salehi, M. Mirzayi, and E. Roudi, "Combinatorial optimization of resilience engineering and organizational factors in a gas refinery by a unique mathematical programming approach," *Hum. Factors Ergon. Manuf.*, vol. 27, no. 1, pp. 53–65, Jan. 2017.
- [22] G. A. Shirali, M. Shekari, and K. A. Angali, "Quantitative assessment of resilience safety culture using principal components analysis and numerical taxonomy: A case study in a petrochemical plant," *J. Loss Prev. Process Ind.*, vol. 40, pp. 277–284, Mar. 2016.
- [23] A. Azadeh, M. Sheikhalishahi, and M. Koushan, "An integrated fuzzy DEA-Fuzzy simulation approach for optimization of operator allocation with learning effects in multi products CMS," *Appl. Math. Model.*, vol. 37, no. 24, pp. 9922–9933, Dec. 2013.
- [24] S. Saati M, A. Memariani, and G. R. Jahanshahloo, "Efficiency analysis and ranking of DMUs with fuzzy data," *Fuzzy Optim. Decis. Mak.*, vol. 1, no. 3, pp. 255–267, 2002.
- [25] L. A. Zadeh, "A Note on Z-numbers," *Inf. Sci. (Ny.)*, vol. 181, no. 14, pp. 2923–2932, Jul. 2011.
- [26] B. Kang, D. Wei, Y. Li, and Y. Deng, "A Method of Converting Z-number to Classical Fuzzy Number," *J. Inf. Comput. Sci.*, vol. 9, no. 3, pp. 703–709, 2012.
- [27] A. Azadeh and R. Kokabi, "Z-number DEA: A new possibilistic DEA in the context of Z-numbers," *Adv. Eng. Informatics*, vol. 30, no. 3, pp. 604–617, 2016.

Computer Science & Systems

CS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to more technical aspects of computer science and related disciplines. The CSNS area spans themes ranging from hardware issues close to the discipline of computer engineering via software issues tackled by the theory and applications of computer science and to communications issues of interest to distributed and network systems. Events that constitute CSNS are:

- BEDA'18—1st International Workshop on Biomedical & Health Engineering and Data Analysis
- CANA'18—11th Computer Aspects of Numerical Algorithms
- C&SS'18 - 5th International Conference on Cryptography and Security Systems
- CPORA'18—3rd Workshop on Constraint Programming

and Operation Research Applications

- LTA'18—3rd International Workshop on Language Technologies and Applications
- MMAP'18—11th International Symposium on Multimedia Applications and Processing

AREA SUPERVISORY COMMITTEE

- Burdescu, Dumitru Dan, MMAP'18
- Damasevicius, Robertas, LTA'18
- Ge, Mouzhi, DaSCA'18
- Janicki, Artur, BigDAISy'18
- Królak, Aleksandra, BEDA'18
- Ksiezopolski, Bogdan, C&SS'18
- Paprzycki, Marcin, 4A'18
- Ristov, Sashko, WSC'18
- Sitek, Pawel, CPORA'18
- Stpiczyński, Przemysław, CANA'18

1st International Workshop on Biomedical & Health Engineering and Data Analysis

IN the recent years, technology has had an accelerating impact on the field of medicine and healthcare. We could observe, in particular, a proliferation of personal health monitoring devices and wearables, such as activity bracelets. These technologies supported by newest developments in data analysis, such as big data and deep learning, have a substantial impact on daily life of many people. They increase awareness of daily activities, help improve sport performance, and can lead to early detection of certain diseases.

The workshop on Biomedical & Health Engineering and Data Analysis—BEDA'2018—provides an open forum for researchers in domains of Biomedical Engineering, Health Technologies, Personal Monitoring Devices, and Data Analysis to communicate high-quality and timely research results. The primary focus is on practical applications, but highly relevant theoretical papers are also of interest.

TOPICS

- Biomedical signal processing;
- Biomedical imaging and image processing;
- Biosensors and bioinstrumentation;
- Neural engineering, neuromuscular systems and rehabilitation engineering;
- Wearable biomedical sensors and systems;
- Health informatics and technology, e-health and telemedicine;
- Biomedical systems management;
- Personal health monitoring devices and wearables;
- Application studies of biomedical and health technologies.

EVENT CHAIRS

- **Królak, Aleksandra**, Łódź University of Technology, Division of Medical Electronics, Poland

- **Wiktorski, Tomasz**, University of Stavanger, Faculty of Science and Technology, Department of Electrical and Computer Engineering

PROGRAM COMMITTEE

- **Agrawal, Bikash**, DNV GL
- **Augustyniak, Piotr**, AGH University of Science and Technology
- **Bajcsy, Peter**, National Institute of Standards and Technology
- **Byambajargal, Byambayav**
- **Caraiman, Simona**, University of Iasi
- **Chakravorty, Antorweep**, University of Stavanger
- **Eftestøl, Trygve**, University of Stavanger, Faculty of Science and Technology, Department of Electrical and Computer Engineering
- **Lundervold, Arvid**, University of Bergen
- **Materka, Andrzej**, Lodz University of Technology
- **Moldoveanu, Alin**, University POLITEHNICA of Bucharest
- **Pissaloux, Edwige**, Université Pierre et Marie CURIE
- **Strumiłło, Paweł**, Lodz University of Technology
- **Strzelecki, Michal**, Lodz University of Technology, Poland
- **Szczypiński, Piotr**, Lodz University of Technology
- **Tadeusiewicz, Ryszard**, AGH University of Science and Technology, Poland
- **Torbicz, Władysław**, Polish Academy of Sciences
- **Unnþórsson, Rúnar**, University of Reykjavik
- **Velázquez, Ramiro**, Universidad Panamericana
- **Vinhais, Carlos**, Instituto Superior de Engenharia do Porto, Portugal

Assistive Smart, Structured 3D Environmental Information for the Visually Impaired and Blind: Leveraging the INSPEX Concept

S. Leseq, O. Debicki, L. Ouvry, F. Birot, L. Sevrin, S. Buckley, C. Jackson, J. Barrett, A. McGibney,
C. Fabre, N. Mareau, J. Foucault, GoSense, SensL, S. Rea, D. Rojas
CEA, LETI, Minatec Campus, Lyon, France, Ireland, Cork Institute of Technology
F-38054 Grenoble Cedex, France, Cork, Ireland
Email: suzanne.lesecq@cea.fr

R. Banach, J. Razavi, M. Correvon, G. Dudnik, J.-M. Van Gyseghem, J. Herveg, N. Grandjean, F. Thiry
University of Manchester, CSEM SA, University of Namur
Manchester, U.K., 2002 Neuchatel, Switzerland, Namur, Belgium

C. O’Murchu, A. Mathewson, R. O’Keeffe, A. di Matteo, V. Di Palma, F. Quaglia, G. Villa
Tyndall National Institute, STMICROELECTRONICS SRL
Cork, Ireland, 80022 Arzano, Naples, Italy

Abstract—Inspired by the abilities of contemporary autonomous vehicles to navigate with a high degree of effectiveness, the INSPEX Project¹ seeks to minaturise the sensing and processing technology involved, to produce devices which can help navigate within an environment in a smart way, enabling their use in a wide variety of applications. The project is focused on producing an advanced prototype for a device which can be attached to a VIB person’s white cane, and which, through the use of a variety of minaturised sensors, and of the processing of their data via sophisticated algorithms, can offer the user a richer palette of information about the environment than the use of the unadorned white cane alone could furnish. The various strands contributing to the project are overviewed, and the prospects for further enhancements are contemplated.

I. INTRODUCTION

SIGNIFICANT visual impairment is highly disabling for any person since a huge proportion of the information people get from the world as a whole comes to them through the visual route. One of the most debilitating aspects of this is the impact on the ability of a visually impaired or blind (VIB) person to navigate unaided through the outdoor environment. Many blind people find themselves confined to their home, simply because venturing outside poses a challenge that they find too daunting.

One way of significantly alleviating this situation is to give the VIB person a guide dog. Guide dogs can be liberating for a VIB person to an almost miraculous degree [1]. A properly

trained dog’s intelligence can process the visual information it sees in the outside world to a level that compares well with that of a human — at least when it comes to recognising dangerous situations and seeing obstacles that should be avoided by its VIB handler. A good guide dog literally gives its VIB handler a new lease of life.

Unfortunately, good guide dogs are hard to come by — demand far outstrips supply. A good guide dog must first be born of the right breed, and display a suitable temperament [2]. Then it must undergo extensive training as it grows from a puppy to an adult dog. Then it must be placed with a VIB owner who must be in a position to look after it, and the two of them must get along unproblematically [1]. None of this is easy, and it costs a lot per guide dog successfully placed.

Advancing technology offers some relief from this unhappy situation. The inspiration comes from autonomous vehicles [3], [4]. These days it is no novelty to see in the news stories of autonomous cars of various kinds navigating with apparent effectiveness through the streets of some city (but now, this can also have tragic consequences [5]). We are told that self-driving cars (possessing a number of degrees of autonomy, graded from 0 (not autonomous at all) to 5 (fully autonomous and dependable in all situations)), are on the horizon. Such vehicles get their navigational capabilities by deploying a large number of sensors, and processing all the data that they produce via sophisticated algorithms whose goal is to produce a navigation strategy for the vehicle that human beings would regard as being appropriate for the circumstances.

The growing familiarity of such a technological approach, combined with the increasing minaturisation of the needed hardware components opens the door to a much wider appli-

¹The INSPEX project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 730953. The work was also supported in part by the Swiss Secretariat for Education, Research and Innovation (SERI) under Grant 16.0136 730953. We thank them for their support. 

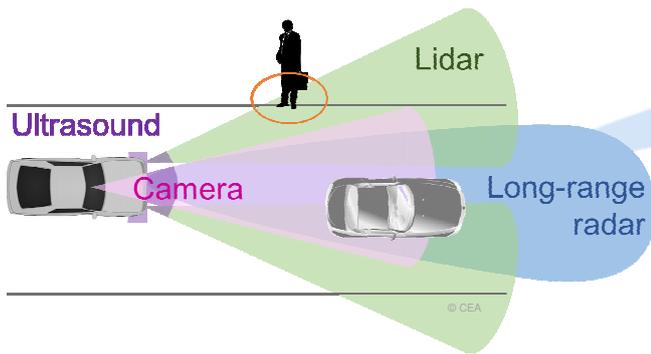


Fig. 1. Sensor deployment in an autonomous vehicle application.

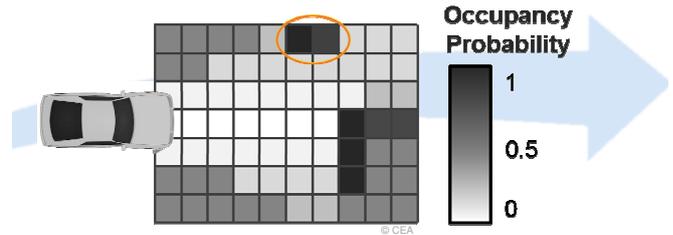


Fig. 2. Turning sensor data into an occupation grid.

cation of the underlying ideas. Autonomous vehicles have the luxury of being able to support an almost arbitrary collection of sensors and processors for this purpose, because of their default size, weight and on board power supply capabilities. Many other applications, however, do not, but that does not make such applications any less desirable.

The INSPEX Project aims to produce an advanced prototype of a minaturised navigation aid that is at a wearable scale. The wearability requirement puts pressure on the number size, weight and power requirements that can be accommodated. While falling short of the full capabilities of an autonomous vehicle, the degree of minaturisation available within today's sensors puts a wearable device within reach. While such a device could be used (and in the longer term is fully intended to be used) for a variety of situations, the creation of an aid that will significantly assist VIB persons to navigate outdoors is the INSPEX Project's priority.

The rest of this paper is as follows. In Section II we briefly overview navigation in autonomous vehicles, and its implications. In Section III we look at how the INSPEX Project positions itself in the light of the preceding. Section IV focuses on the VIB use case. In Section V we outline the issues faced by the INSPEX design. Section VI looks at the INSPEX advanced prototype in detail, and Section VII describes the associated physical integration challenges. Section VIII looks at verification, validation and evaluation in INSPEX. Section IX surveys related systems and Section X concludes.

II. NAVIGATION IN AUTONOMOUS VEHICLES

The inspiration for the INSPEX Project comes from the growing capabilities of autonomous vehicles. Conceived in fiction over three decades ago, autonomous vehicles started to become a serious reality within the last decade, especially the last five years. There are legion societal issues surrounding true vehicle autonomy, from attribution of liability when something really bad happens, to massive job losses due to a vanishing need for drivers. Our interest, though, is in the technology.

Fig. 1 shows a schematic of an autonomous vehicle system. The car on the road is equipped with LiDAR, with RADAR, with ultrasound (US) as well as a camera and vision system;

not shown in the figure, there is also an IMU. On board the vehicle, there will be a GPS system together with a mapping application.

In a real driving situation, the data pouring from the sensors is gathered, timestamped, and then sent to a data fusion application that calculates, using Bayesian estimation, an *occupancy grid* [6], [7]. This divides the 3D space in front of the vehicle (and to the extent necessary, the sides and back too) into cubical sections, for each of which, an estimate is calculated of the probability of there being an obstacle of some sort present in that cube. Fig. 2 gives an idea of this, in which the car in front of the instrumented car of Fig. 1 is detected by the long range RADAR and the person standing at the side of the road is detected by the LiDAR sweeping to the side. The occupancy grid in Fig. 2 shows the higher probability of there being an obstacle in these positions by the denser shading in the relevant squares.

As noted earlier, the large size, weight and power of a vehicle permits a full suite of useful sensors to be deployed. In a minaturised scenario, this is no longer possible, and the main casualty is the vision system. It is not that cameras need to be large or heavy these days, but the processing required to make useful sense of a visual image would overwhelm the capacity of any battery light enough to be carried (and the capability of any computing equipment not connected to the cloud).

III. THE INSPEX PROJECT AND ITS USE CASES

Taking on board that the contemporary trend in sensors of all kinds is towards ever smaller ever more low power versions of the basic capability of the sensor, INSPEX intercepts this trend to reappraise the autonomous vehicle obstacle detection and navigation architecture. The outcome of this is the concept of a small and lightweight device, minimising the resources needed for efficient occupancy grid estimation [8], that offers the potential to enhance the way that navigational and positioning challenges are addressed in a wide variety of application areas.

Fig. 3 gives an indication of the range of applications whose working might benefit from the incorporation of INSPEX technology. The figure is divided into several sections.

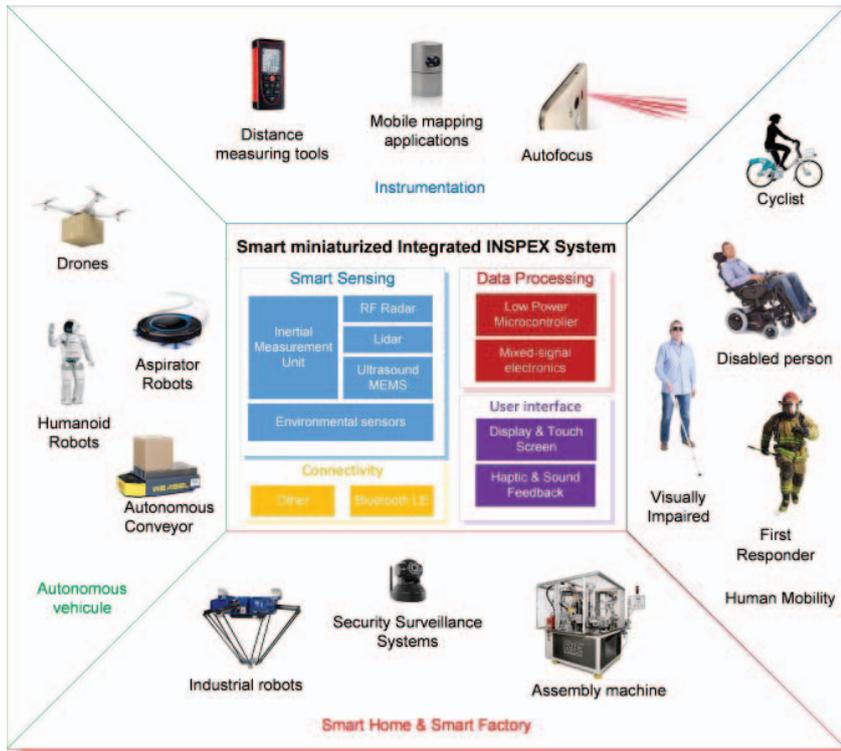


Fig. 3. A selection of potential INSPEX use cases.

On the right we see schematics for human centred applications. We see the VIB use case which forms the focus of the initial INSPEX development. Accompanying this are other use cases. The first responder use case covers situations like that of firefighters, who regularly need to attempt rescues and other complex retrieval operations in enclosed smoke filled (and thus opaque) environments and thus can benefit from non-visual cues about obstacles around them. The severely disabled who have obstacles to process the visual information from their surroundings can also benefit from INSPEX. There are also use cases for the non-disabled, such as cyclists and pedestrians. Such use cases arise, particularly in various highly polluted mega-cities, where, on a regular basis, smog levels reach such a pitch that visibility is seriously restricted.

At the top of Fig. 3 we have some use cases connected with instrumentation. Modern distance measuring tools typically make use of a single laser beam whose reflection is processed to come up with the numerical result. This strategy works well when the target is a hard, flat surface. For more uneven, and textured surfaces, the reading given (if any) may be subject to errors. An INSPEX enhancement to such instruments can yield better error detection and could perform better in a wider variety of situations. INSPEX could also improve to working of mobile mapping applications. Modern 3D environment capture instruments already use LiDARs to accurately measure the characteristics of their environment. Integrating more sensors into such instruments can improve their performance by better

registering elements of the environment that are transparent to LiDAR. Similar remarks apply to autofocus applications.

These days, camera systems (typically in leading edge phones) employ increasingly sophisticated algorithms to distinguish foreground from background, to make up for varying lighting conditions, and generally to compensate for the user's lack of expertise in photography. Bringing in the augmented sensing capabilities of INSPEX opens the way to even more sophisticated approaches in this field.

At the bottom of Fig. 3 we have some use cases mainly concerned with navigation inside large, enclosed environments. Currently, highly automated factories can feature assembly lines consisting of hundreds of robots, each performing a specific task, within tightly constrained parameters. Such factories have to be planned and laid out with a great deal of precision because the industrial robots involved have limited intelligence. The proliferation of big data

techniques generates an impetus to increase this robotic smartness as rapidly as possible, so that the detail needed in the planning phase can be reduced, and the robots involved can be more rapidly reconfigured to address the tasks spawned by the changing business environment. Adding INSPEX capabilities to the mix enhances the ability of the robots to accurately orient themselves relative to other significant elements of their working environment. And while the scenario just outlined focuses on fixed robots, the issues to be addressed become even more acute if the robots are mobile. Mobile robots need to accurately orient themselves within their environments, particularly so that they can avoid harm to humans who may be working in their vicinity. Also, security surveillance systems, which traditionally rely on infra-red sensors, can benefit from the enhanced precision that INSPEX can bring.

On the left of Fig. 3 we see some use cases in the autonomous *small* vehicle domain. When a vehicle is small, the effectively unrestricted tolerance of high weight, size and power consumption that large road vehicles can afford, disappears. Small airborne drones have demands of this kind that are every bit as exacting as those that we have in the VIB use case. As the use of drones increases, especially within cities, their need to accurately navigate within the urban environment rises significantly. Drones will increasingly need to manoeuvre within quite tightly constrained urban spaces, without colliding either with the fixed environment, or with each other. INSPEX capabilities can significantly address the sensing requirements of this use case. Humanoid robots have similar requirements since they are quite severely constrained regarding (especially) size and power consumption. INSPEX

capabilities will increasingly become indispensable as such robots increasingly become autonomous and capable of independent action. Another use case highlighted in the figure is that of autonomous domestic robotic assistants. A typical floor cleaning robot will, these days, use a fairly hit and miss approach to finding parts of the floor to clean. Gradually, it learns where obstacles are, and builds an approximate map to improve its navigation subsequently. This works well if the furniture remains static. But if the user is inclined to move things around to repurpose the room in different ways, then the robot will remain in the early learning phase as its prior experience is frequently rendered unreliable. More accurate navigation capabilities acquired from equipment such as INSPEX can contribute to shortening the learning curve in these circumstances.

IV. THE VIB USE CASE CHALLENGE

The main focus of the INSPEX Project is the design and production of an advanced prototype obstacle detection and navigation system to assist visually impaired and blind users. According to the World Health Organisation, the number of VIB persons numbers 285 million and, due to an aging global population along with the debilitating health conditions that greater age brings, this number will double in twenty years or so [9]. The wide variety of conditions that a person may encounter in the outdoor environment makes the design of a system that gives really useful feedback to users in a convenient and informative way into a highly nontrivial challenge. The primary objective of INSPEX is to produce a clip-on device (together with the support facilities that it needs) that can add significant functionality to the principal navigational aid used by VIB persons, namely the white cane.

Obstacles to the VIB person's smooth and safe progress come in many forms. They include actual solid objects that impede the user — these may be found at many levels, from the ground, all the way up to head height. This variety of positions at which obstacles may be located poses a major problem for detection technologies, as an approach that detects at one position may be oblivious to obstacles elsewhere. Somehow, the whole range of the user's 3D space must be monitored and the user suitably informed.

Obstacles on the ground are dealt with quite effectively using the white cane. The experienced VIB user can gain a lot of information, especially along familiar routes, from the minor variations in the surface texture that they sense with their white cane. The specially textured tiling that has been introduced along urban pavements, particularly at junctions, and along the edges of pavements and platforms for light and heavy rail, are a deliberate attempt to increase the informativeness of the surface for the disabled user, and for the VIB community in particular [10].

In urban environments, especially in changeable ones, collisions with obstacles at chest or head height are a significant risk factor for the VIB community. VIB persons may wear headgear, even if for sartorial reasons they might prefer not to, to offer at least some cushioning in case of unanticipated

collisions. Such strategies offer, at best, only partial defense against the risk, and injuries sustained by VIB persons are a *de facto* unavoidable feature of a VIB individual's efforts to participate in normal life [11]. In theory, VIB persons could choose to wear crash helmets and similar gear, but such items are felt to be extremely stigmatising, and so are eschewed by the VIB community.

We have been speaking about solid obstacles with which VIB persons might collide, but these are not the only hazards to worry about. Danger for the VIB person arises from approaching empty holes in the surface and holes filled with water (or other materials, whether liquid or not). Staircases comprise a vital element of the contemporary urban landscape, and constitute a particular risk factor for the VIB, especially when the staircase is a descending set. (It is worth observing that escalators, while being similar to stairs, are less risky for VIB persons, since they emit a typical 'escalator rumble' which warns the VIB person of their proximity.)

Thus a wide variety of sensing capabilities is needed to address all of these situations. Moreover, different meteorological conditions can significantly affect what a sensing instrument can acquire. Each different kind of precipitation has its own transparency/opacity characteristics, impacting the fidelity of the information that can be inferred from different sensors under different conditions.

V. INSPEX VIB DESIGN ISSUES

It is very tempting, when developing a technology to meet a user need, for the technologists to get completely absorbed in the fine details of the technical aspects and to drift away from involvement in the needs of the prospective user community. This simply decreases the degree to which the end result actually meets the original need.

To forestall this eventuality in INSPEX, an early phase of the design process was focused on user needs. Significant effort was expended to engage with associations for the VIB community to elicit the requirements that improved mobility aids ought to address. Following this engagement, a collection of personas were constructed in order that typical activities for such personas could be identified, to ensure that the eventual prototype that would be developed could be validated against these identified activities.

With this preparatory foundation addressed, the design phase could proceed in earnest. The overall architecture of the INSPEX system consists of three modules. There is the hardware module that attaches to the VIB person's white cane. This module communicates via Bluetooth with a software component running on a smartphone belonging to the user. The software component is able to use the phone's capabilities to communicate with the outside world, and to integrate information received from there with information gained locally, in order to give the VIB user a richer and more informative experience. Finally, there is a pair of extra-auricular headphones, connected to the smartphone module again via Bluetooth, that is able to paint a sound picture for the VIB user about their environment that is intended to be as vivid



Fig. 4. The complete INSPEX system for the VIB use case.

and informative as possible. Fig. 4 shows these components in digramatic form.

The account just given makes it clear that INSPEX belongs very much in the *Internet of Things* family of systems. An INSPEX system is ready to receive information made available by the environment, in order to assist its user to gain a more accurate picture of where, in relation to the rest of the environment, he or she might be. This potentially raises various ethical concerns. These days many smart apps are designed to divulge data about their users, whether to the app manufacturer or to connected nodes in the wider environment, and it is the users' (assumed informed) choice about whether or not they are prepared to have such information divulged. By contrast, from a legal perspective, the VIB use case of INSPEX is generally regarded as a Class 1 medical device [12], [13]. Because of the medical connection, by default, a much larger quantity of regulation comes to bear on the device (although the regulatory situation may change in future). Moreover, the details of the regulations concerned vary from jurisdiction to jurisdiction, so ensuring that INSPEX conforms to all required regulation forms a significant strand of the project (handled by partner Univ. Namur). By default, INSPEX implements a *privacy by design* policy. No data (that is intelligible outside the INSPEX system) is ever made available to the environment.

Fig. 5 illustrates a prototype configuration of a VIB user deploying a white cane augmented with the kind of sensing abilities envisaged for the INSPEX white cane hardware clip-on module. In fact the figure shows an early conceptualisation of the clip-on module, when having an additional module near the bottom of the white cane, to more precisely monitor conditions near the ground, was contemplated. In actual fact, consideration of the user feedback gained from the community and persona studies earlier in the process, quickly convinced the project consortium that the leverage around the

white cane's point of support from even a very light component attached near the end of the white cane would rapidly generate an unacceptable strain on the user's wrist (not to mention the added complexity, from the user's perspective, of having to deal with one more device, most likely with its own charging regime), rendering such a design unacceptable, even if it led to some engineering advantages. Therefore, the current design of the hardware clip-on module is a single unit attached near the point of support of the white cane by the user.

VI. THE INSPEX ADVANCED PROTOTYPE SYSTEM

In the light of the many and varied environmental circumstances that a VIB person may meet outdoors, the hardware clip-on module features a collection of sensors, that are intended to harvest the maximum amount of useful data in the widest possible range of conditions, bearing in mind the limitations of size, weight and power consumption. In this section, we overview the various components that comprise the INSPEX advanced prototype.

The heart of the INSPEX sensing system consists of four main sensors. There is a short range LIDAR, a long range LIDAR, a MEMS ultrasound sensor and a ultrawideband RADAR. These are supplemented with an inertial measurement unit and appropriate computation and communication facilities. The collection of sensors that have been incorporated into INSPEX complement each other in important ways, making up for each others' deficiencies under different circumstances in ways that genuinely add value [14].

For example, ultrasound sensors have a limited range (typically not exceeding 3m). While a 3m range is largely adequate for the VIB use case, it is insufficient for many other use cases identified earlier, such as drones, autofocus, and surveillance systems.

Sensors based on lasers can see much further, but have other deficiencies, such as their relatively poor performance under conditions in which the ambient light shares frequencies or harmonics with the sensing system. They also perform poorly in detecting objects that are transparent to their frequencies; likewise highly reflective surfaces which do not scatter widely enough to send enough energy back along the incident beam.

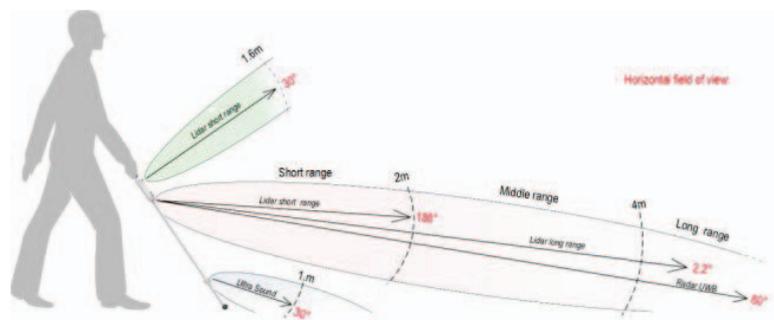


Fig. 5. Prototype INSPEX VIB white cane use case.

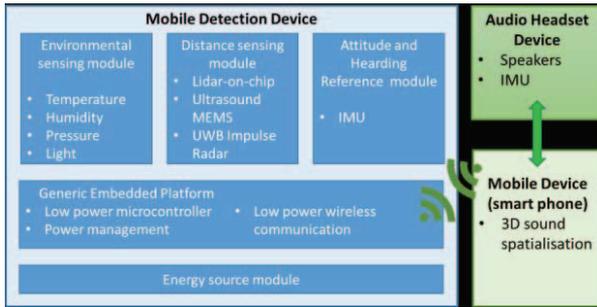


Fig. 6. The architecture of the INSPEX system.

Radio frequency radar, which utilises a different part of the electromagnetic spectrum than lasers, can perform much better when lasers are impeded by poor weather conditions such as rain, sleet, snow, hail. RF radar is also unimpeded by small obstacles, due to its wavelength. These response characteristics are also very different from the mechanical response of obstacles to ultrasound. It is thus clear that three technologies mentioned do offer the good complementarity claimed.

A. Overall INSPEX Architecture

Fig. 6 shows all the components of INSPEX. In the main Mobile Detection Device we see not only the distance sensing module whose elements we have just discussed, but also other modules. There is an Environmental Sensing Module. This provides information about the ambient circumstances of use of INSPEX to improve the decision making capability of the overall system about how to best configure itself for optimal performance. Vital to effective working is the Attitude and Heading Reference Module. This contains an Inertial Measurement Unit, which regularly feeds information about the physical orientation in real space of the whole Mobile Detection Device to the system's decision making capability. This is indispensable as all the distance measurements obtained by the distance sensing devices, must be correlated to the direction in 3D space in which they were made, in order for them to be useful in constructing a 3D occupancy grid. Energy is supplied to the whole Mobile Detection Device by the Energy Source Module, which is an interchangeable battery pack that can be quickly swapped for a spare when power is low, and which offers rapid recharge characteristics. The Generic Embedded Platform coordinates and controls the activities of the other modules in the Mobile Detection Device. As well as this it takes care of communication with what is referred to as the Mobile Device (Smart Phone) in Fig. 6.

A separate major component is the Audio Headset. This features a pair of extra-auricular earbuds to project the audio signal from INSPEX into the user's ears. It also contains another Inertial Measurement Unit, so that the orientation of the user's head may be inferred, in order that the calculated INSPEX audio signal is perceived to be oriented stably with respect to a 3D space inertial frame.

The final major component of the system is the smartphone. This is connected to both the Mobile Detection Device and the Audio Headset via Bluetooth. It handles the IoT element of INSPEX's functionality which listens out for IoT beacons in the environment, information which is integrated into the audio signal presented to the user. It also correlates the 3D information of both Mobile Detection Device and Audio Headset to ensure the 3D space inertial frame stability required by the user.

B. The INSPEX Development Trajectory

The INSPEX development trajectory foresees the achievement of the final advanced prototype through a series of milestones, one per year, each producing a working version of the system that partly realises the project goals. Thus, the first year of the project produced P0, the first (non-minaturised) prototype. The second year produces P1, featuring the first stage of minaturisation and integration. The third year culminates in P2, the final prototype, in which minaturisation and integration reach their target level.

In the subsections that follow, we examine the various elements of INSPEX that have been introduced earlier in a little more detail. We start with the various sensors, and we illustrate the description of these with pictures of their early P0 prototypes. These were the only ones that had been manufactured and shown to be working at the time of writing of this paper.

C. The Ultrawideband RADAR (CEA)

The INSPEX RADAR sensor, developed by partner CEA, is an ultrawideband radar designed to detect movement in obstacles located up to 20m away. Refreshing at up to 100Hz will make it capable of high performance tracking of moving targets. This is facilitated by an aperture of 60°, giving the radar a reasonably wide viewing angle.

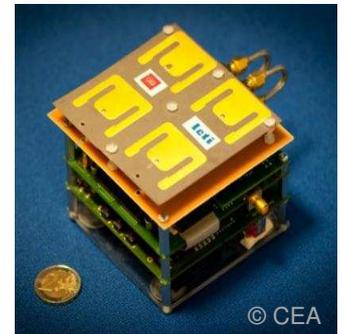


Fig. 7. The UWB Radar (CEA).

Fig. 7 shows an early prototype of the radar. Minaturisation of such radars is not readily seen in the electronic component market, limiting the prospects for INSPEX. As INSPEX develops towards its final version, the first prototype will be minaturised to the extent that is feasible, with the circuit boards visible in the figure reduced in size. The antenna elements allow little scope for size or weight reduction.

D. The MEMS Ultrasound (ST Microelectronics)

The INSPEX ultrasound sensor, developed by partner ST Microelectronics, is a microelectromechanical system, or MEMS sensor. Of the two types of these that are available,

capacitance based and piezoelectric effect based, the piezoelectric kind offers much lower power requirements, so is preferred for use in INSPEX. Fig. 8 shows an early prototype implementation for INSPEX. The size of the circuit board will decrease significantly in the later stages of the INSPEX project.



Fig. 8. The MEMS ultrasound sensor (ST Microelectronics).

in a large proportion of modern cars.

E. The Long Range LiDAR (Tyndall, SensL)



Fig. 9. The long range LiDAR (Tyndall, SensL).

The Long Range LiDAR sensor for INSPEX, developed by project partners Tyndall and SensL, is a narrow beam (2.2°) LiDAR with a range of around 20m or more. An early prototype appears in Fig. 9. The long range is useful for recognising distant hazards, however the narrow beam angle militates against utility. To mitigate this disadvantage, an array version, utilising 64 channels is being developed for INSPEX. This will effectively scan (a portion of) a plane passing through the sensor, registering the distance away of obstacles encountered by the beams that radiate from the sensor in directions embedded in the plane.

Given the natural orientation of the Mobile Detection Device on the user's white cane, the plane will be oriented vertically, and so the user will gain a picture of obstacles at larger distances by sweeping the white cane from side to side, in a motion that departs little from a VIB person's natural use of a non-INSPEX-enhanced white cane. Further development of the sensor will be principally focused on reducing the size of the circuit board.

F. The Short Range LiDAR (CSEM, CEA)

The Short Range LiDAR sensor for INSPEX, is developed by project partners CSEM and CEA. It is based on an off-the-shelf component which is shown in Fig. 10. This LiDAR component has a larger field of view (around 30°) and a range of around 2m.

Ultrasound sensors have been used in diverse applications for some time now, for example, non-destructive testing, flow metering and medical imaging. The application perhaps most familiar to the man in the street is the collision warning system, to be found in various guises, such as parking assistance systems,

As in the long range LiDAR case, the basic OTS component will be integrated into a second generation sensor capable of looking at 9 independent measurement points within its field of view. Also as for the long range LiDAR case, development effort for the short range case will miniaturise the electronics, leading to a much smaller circuit board in the final prototype.

Fig. 11 shows an early version of the short range LiDAR integrated into a configuration that can sweep a plane, in the manner described for the long range LiDAR case.

G. The IMU

The Inertial Measurement Unit in the Mobile Detection Device of INSPEX is an off-the-shelf component, these devices being small enough these days to use without adaptation (this, in turn, being due to their mass use in mobile phones). Its job is to regularly supply information about its orientation and displacement in Euclidean space (and thus about the orientation and displacement of the Mobile Detection Device itself, and therefore about the orientation and displacement of the sensors contained within it) to the Generic Embedded Platform. This is needed so that the distance information coming from the other sensors can be correctly related to the current state and orientation of the occupation grid.

H. Software and Firmware in the Generic Embedded Platform

The Generic Embedded Platform within the INSPEX Mobile Detection Device has the responsibility of appropriately integrating all the sources of information provided by the family of sensors embedded in the Mobile Detection Device, and communicating suitably with the INSPEX Smartphone Application. Considering the variety of the information received, the integration of all this information is a non-trivial task, and a considerable quantity of sophisticated software is needed to accomplish an optimal result.

A major portion of the job is the receipt and husbanding of the raw data coming from the sensors. These all work at their own rate, modulated by the commands sent to them by the GEP. The data has to be received and timestamped so that fresh data may be given priority over older stale data. The raw



Fig. 10. The short range LiDAR COTS element.

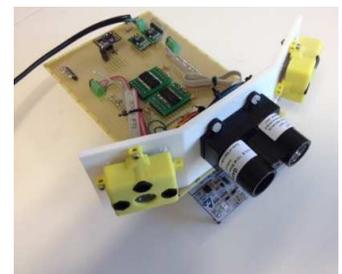


Fig. 11. The short range LiDAR in wide angle configuration (CEA, CSEM).

distance data has also to be correlated with data from the IMU so that the orientation in space of the distance measurement may be discerned.

Once the raw data has been correctly marshaled, it is passed to the occupation grid computation, a proprietary component supplied by partner CEA. The CEA occupation grid computation is distinguished by its significantly greater efficiency [8], compared with standard Bayesian estimation algorithms, with their reliance on heavy duty floating point calculation [6], [7], which has a significant impact on power consumption in a portable lightweight system such as INSPEX.

Once the occupation grid has been derived, it is transmitted via Bluetooth to the INSPEX smartphone application. The granularity of the occupation grid is heavily constrained by the bandwidth available for this communication, and this places a fundamental limit on the detail about the environment that can be communicated to the user (although, see the discussion about this in Section VI-K).

I. Power Management in the Mobile Detection Device

The Mobile Detection Device contains a significant number of hardware components. Albeit that these are all ultimately minaturised low power components, they all still require power. As invariably happens in projects of this kind, each component in fact consumes more power than initial estimates stated, so significant effort has to be invested to reduce this, and to deploy every means possible to optimise power use.²

In INSPEX, a sophisticated power management strategy is utilised to eke out the energy made available by the energy source module to the greatest effect. A formal modelling approach is used to organise the various factors that have an impact on the consumption of energy into a consistent approach, consisting of three layers.

At the simplest level there is the correctness layer, which ensures that no hardware element is left running when it should not be doing so, and handles similar concerns. Next is the quantitative layer, which enhances the preceding with quantitative information, enabling estimates to be derived about power consumption and system uptime under different circumstances. Lastly comes the policy layer, which imposes constraints that take environmental and other considerations into account, enabling best use of remaining power according to higher level, user determined goals.

J. The Audio Headset

The INSPEX audio headset performs two important roles within the functioning of the whole system. It contains a pair of extra-auricular earbuds whose job it is to project the audio signal received via Bluetooth from the INSPEX smartphone application into the user's ears. The earbuds are extra-auricular in order that the user not be prevented from hearing sound from the environment itself. Many VIB persons report that they can 'hear' all sorts of elements within their environment (e.g. doors open or closed within the indoor environment, or

walls and lampposts within the outdoor environment) unaided, so it is vitally important that they should not be deprived of this ability when using INSPEX, or else the takeup of the INSPEX VIB system by its target market will be very poor.

The INSPEX audio headset also contains another IMU. Thus the second job done by the audio headset is to send regular information about the orientation and displacement in Euclidean space of the audio headset to the INSPEX smartphone application. This enables the binaural audio signal received from the smartphone to consist of sounds that are computed to have a stable point of origin in Euclidean space, despite the movement of the user's head.

K. The INSPEX Smartphone Application

The INSPEX Smartphone Application, developed by partner CIT, binds the functionality of the INSPEX modules discussed above into a consistent whole. One major task is the integrated handling of all the 3D information coming from different parts of the system. The mobile detection device is physically independent from the audio headset, so each is independently situated in Euclidean space. The information coming from the two IMUs in these modules has to be integrated so that a single consistent picture of the INSPEX system's position in 3D space can be determined. The calculations for this, though not trivial, are relatively familiar from the robotics world [15], [16].

Another major responsibility of the smartphone application is context awareness. In the growing world of IoT, especially as realised in the growing number of *smart cities*, the environment is capable of being, of itself, informative in the cyberspace domain. A smartphone will be capable of picking up this information from *smart beacons* as its user moves through such a smart city. Although a significant proportion of the smart beacons implemented in smart cities will have a commercial dimension, a proportion will be intended to act as aids to disabled users, especially the VIB community. The INSPEX smartphone application will be able to receive the signals from these, and to integrate it into the audio information transmitted to the user.

A final major responsibility of the smartphone application will be synthesising the sounds to be transmitted to the audio headset via Bluetooth. Here there is a significant challenge. The visual domain is not only very rich due to the high resolution available in visible frequencies, but is also three dimensional. By contrast, the audible domain, although also having a spectral range, is much more compressed than the visible spectrum, and sound needs a temporal element to be heard, whereas a visual impression can be gained almost instantaneously. How then to design audio feedback to users that accommodates the following conflicting requirements:

- the varying capability to overlay audio information onto ambient sound from the environment according to the ambient sound level and characteristics,
- the user's desire to hear the natural environment clearly,
- the user's desire not to be overloaded with audio information that is too complicated to absorb,

²Exactly the same remarks invariably apply to the overall weight of the integrated system.

- the user's desire to receive rich audio information, especially about the most important hazards in the environment, whatever they might be.

INSPEX partner GoSense is responsible for this constituent of the design, which is tackled via an *Augmented Reality Audio* approach. While good progress is anticipated with this approach, one can imagine an ongoing artificial intelligence challenge, stretching beyond the project's lifetime, to discover how to best tailor the audio feedback, in the face of the conflicting requirements listed, for different types of users, who can have very different preferences regarding what works best for them.

VII. PHYSICAL INTEGRATION

One of the major targets of INSPEX is systems integration to achieve a small volume (100cm³), lightweight (200g) system 'Box' that meets the requirements of the VIB application. Into this Box must be integrated all of the obstacle detection transducers along with their signal processing boards, the main processing board, battery, power management and all cabling, connectors and switches.

Beyond the above, there are many other physical and practical constraints to be taken into account. Thus, each obstacle detection transducer must be orientated to point in the appropriate direction, and EMC issues must be managed properly inside the densely-packed Box. The Box and its contents must be robust against the mechanical impacts and vibrations induced by cane use and impact with obstacles, and the Box must be weather- and dust-proof while still allowing each obstacle detection transducer to transmit and receive through the Box wall. It must be possible for the VIB user to easily attach and detach the Box from the white cane either for off-cane charging or to allow the cane to be used without the INSPEX System if the user is, for instance, using the cane in a very familiar area. And to minimise imbalance and strains that could lead to user fatigue, the Box also be weight-balanced on the cane, with its centre of gravity as close as possible to the long axis of the cane. Finally, the Box should be aesthetically pleasing and easy to clean for the user.

Combined, these requirements represent a major hardware design challenge that must be addressed at every level of the system, combining both bottom-up and top-down design approaches, and many variations on systems integrations concepts, such as shown in Figure 12, to iterate to a whole-system solution. Partner CIT takes care of these aspects of the INSPEX development.

VIII. VERIFICATION, VALIDATION, EVALUATION

A system as complex as our presentation of INSPEX has elucidated, evidently throws up highly non-trivial issues regarding the ensuring of correct behaviour under all operating conditions. This challenge must be approached in a disciplined and structured way, otherwise little assurance is derived.

At the lowest level, individual project partners are responsible for the testing and verification of their individual submodules. Usual best practice in each individual physical

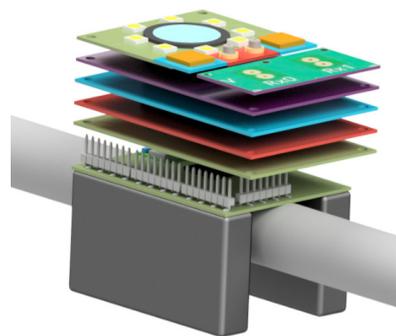


Fig. 12. Prototype INSPEX VIB physical configuration.

domain is applied in each case. At the next level up, the behaviour of the community of sensors must be integrated via the software and firmware of the generic embedded platform. Here, software complexity must be confronted. In this sphere, formal modelling and verification techniques are brought to bear. Since so much of the system depends on the behaviour of physical systems, which, compared with software systems, are considerably less negotiable (in the sense that one might simply change one's mind about what software should do—almost at will—whereas with physical devices, one is restricted to what the laws of nature allow and device engineering will permit), the basic objectives and low level aspects must come from the hardware development. Once those aspects are clear however, the higher level aspects can be modelled using a formal framework such as Event-B [17], [18]. This done, the required consistency of the models can be established, and refinement techniques can progress the models close to code. Partner Univ. Manchester is responsible for this interworking between formal and conventional development techniques. The result of all the activity that has been described can then be confronted with the project verification plan, which was drawn up *a priori* to define correct operation for INSPEX.

Once correct operation has been established, validation against user requirements can be done, to yield an evaluation of the achievements of INSPEX against real user needs. A number of VIB user communities are in communication with the project team, and will contribute to such user evaluations after the technical development has been completed.

IX. RELATED SYSTEMS

The idea of creating an assistive device to help VIB persons to navigate safely, that functioned by translating information gained via wave physics based sensing into aural or tactile information communicable to a VIB person, does not originate with INSPEX. Among existing systems we can mention Smartcane [19], Ultracane [20], Bawa [21], Rango [22], etc.

Aside from these initiatives that are focused on making white canes smart, there are many devices that are deployed via mechanisms such as headwear of one kind or another, clip-on devices that clip on to the user's belt, or devices that can be strapped on to the chest or hung round the neck.

However, the reaction from the VIB community to these proposals has been largely negative, as they are perceived to be stigmatising to a significant degree. And besides these products that primarily originate in the SME sector, the increasing promotion of societal assistance for citizens that suffer from various disabilities spurs an increasing interest from large corporations, e.g. [23].

The key observation about earlier systems is that they typically use a single sensing channel to probe the 3D environment in front of the user. This reduces the task of translating the information obtained into a format comprehensible by the user to the mere indication of a single distance measurement (in the simplest cases). This simplification avoids the complexities and costs of data fusion, at the price of forcing the toleration of any weak elements or ‘blind spots’ that are inherent in the physics of the given sensing channel. INSPEX is one of the earliest VIB assistive devices that is intrinsically multi-channel, although it is seen that more and more such devices are being proposed nowadays.

X. CONCLUSION

In the preceding sections we have introduced the INSPEX Project, and its ambitious retinue of imaginative use cases. The project currently focuses on the VIB use case, and the issues raised by this use case have been elaborated in detail in this paper. When the advanced prototype for the smart enhancement to the white cane has been completed, development will be pursued further towards a fully fledged commercial product. This subsequent phase will seek to further reduce the size, power consumption and weight of the unit, and will endeavour to locate a price point that will render it attractive to the VIB community. It is anticipated that the fruitful exploitation of this INSPEX use case will act as a spur to the development of further commercially exploited use cases from among the wide selection described in Section III.

REFERENCES

- [1] L. Whitmarsh, “The Benefits of Guide Dog Ownership,” *Visual Impairment Research*, vol. 7, pp. 27–42, 2005.

- [2] E. Wilsson and P.-E. Sundgren, “The Use of a Behaviour Test for the Selection of Dogs for Service and Breeding, I and II,” *Applied Animal Behaviour Science*, vol. 53 and 54, pp. 279–295 and 235–241, 1997.
- [3] Z. Qu, *Cooperative Control of Dynamical Systems: Applications to Autonomous Vehicles*. Springer, 2009.
- [4] M. Fausten, “Evolution or Revolution: Architecture of AD Cars,” in *Proc. IEEE ESWEEK*, 2015.
- [5] UBER Self Driving Car Fatality, <https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe>.
- [6] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. MIT Press, 2005.
- [7] Moravec, H. and Elfes, A., “High Resolution Maps from Wide Angle Sonar,” in *Proc. IEEE ICRA*, 1985.
- [8] Dia, R. and Mottin, J. and Rakotavao, T. and Puschini, D. and Leseq, S., “Evaluation of Occupancy Grid Resolution through a Novel Approach for Inverse Sensor Modeling,” in *Proc. IFAC World Congress, FAC-PapersOnLine*, vol. 50, 2017, pp. 13 841–13 847.
- [9] WHO. Visual Impairment and Blindness, <http://www.who.int/mediacentre/factsheets/fs282/en/>.
- [10] T. Rosburg, “Tactile Ground Surface Indicators in Public Places,” in *Human Haptic Perception: Basics and Applications*, Grunwald, Ed. Springer, Birkhauser, 2008.
- [11] R. Mandruchi and S. Kurniawan, “Mobility-Related Accidents Experienced by People with Visual Impairment,” *Insight: Research and Practice in Visual Impairment and Blindness*, 2011.
- [12] V. Theisz, *Medical Device Regulatory Practices: An International Perspective*. Pan Stanford Publishing, 2015.
- [13] Wikipedia: Medical Devices, https://en.wikipedia.org/wiki/Medical_device.
- [14] L. Scalise, V. Primiani, and P. Russo, “Experimental Investigation of Electromagnetic Obstacle Detection for Visually Impaired Users: A Comparison with Ultrasonic Sensing,” *IEEE Trans. on Inst. and Meas.*, vol. 61, pp. 3047–3057, 2012.
- [15] R. Paul, *Robot Manipulators: Mathematics, Programming, and Control*. MIT Press, 1982.
- [16] R. Murray, Z. Li, and S. Sastry, *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.
- [17] J.-R. Abrial, *Modeling in Event-B: System and Software Engineering*. CUP, 2010.
- [18] RODIN Tool, <http://www.event-b.org/> <http://sourceforge.net/projects/rodin-b-sharp/>.
- [19] Smartcane, <https://www.phoenixmedicalsystems.com/assistive-technology/smartcane/>.
- [20] Ultracane, <https://www.ultracane.com/>.
- [21] Bawa, <https://www.bawa.tech/>.
- [22] Rango, <http://www.gosense.com/rango/>.
- [23] Apple VIB Navigation Announcement, <https://appleinsider.com/articles/18/06/28/apple-mulls-system-for-helping-visually-impaired-navigate-environment>.

Development of a mathematical model for electrode systems in rheophthalmography

Anna Kiseleva

Bauman Moscow State Technical
University
Moscow, Russian Federation
Email:kiseleva.anna.a94@gmail.com

Petr Luzhnov

Bauman Moscow State Technical
University
Moscow, Russian Federation
Email:petervl@yandex.ru

Dmitry Shamaev

Bauman Moscow State Technical
University
Moscow, Russian Federation
Email:shamaev.dmitry@yandex.ru

Abstract—The problem of estimating the electrical impedance characteristics was solved using the system of impedance diagnostics of blood circulation with the help of mathematical modeling. In this work, the geometry for mathematical modeling was reconstructed; its basic quantitative characteristics were calculated. The working capacity of the model is verified on the basis of theoretical data. An example was shown by using the model to select the optimal positions of the electrodes for conducting electrical impedance studies in rheophthalmography. As a result, an example of simulation was shown.

Keywords — mathematical modeling, electrical impedance, eye, pulse blood filling.

INTRODUCTION

Mathematical modeling is one of the powerful tools of modern science [1,2]. Unlike experimental studies, the methods of modeling make it possible to change the parameters, characteristics, and properties of the system under investigation in a wide range. Advances in computer technology make it possible to overcome the difficulties of analytical calculations in the study of complex models, which allows to obtain quantitative characteristics [3].

Modeling has become widespread in the field of biomedical technology [4,5,6]. This is due to the fact that in the development of complex biomedical systems, mathematical modeling is the only way to assess the quality of their functioning in advance. Moreover, it allows to simplify the design phase synthesis and analysis processing of the data, the implementation of which requires digital computing devices, as well as optimizing control algorithms and external interaction with other systems.

In our paper, the task of estimating the electrical impedance characteristics is solved using the rheophthalmography system [7,8] with mathematical modeling. Rheophthalmography is an electrical impedance method for the studying pulse oscillations in the blood filling of the eye vessels, based on the graphical recording of changes in the total electrical resistance of tissues. Optimal positions for electrodes have been chosen and a model has been developed for further verification of the results obtained.

MATERIALS AND METHODS

A. Statement of the problem

It is necessary to determine the boundary conditions of the solving problem in the compilation of a mathematical model of any system. For the purpose of carrying out a numerical experiment, it is necessary to compile an algorithm for the numerical method for solving equations which is a mathematical model of the process under study. As a result, the model is refined itself and the algorithm is corrected [9].

Methods of mathematical modeling techniques are dynamic systems theory. Tools - differential and difference equations, methods of qualitative theory of differential equations, computer simulation [10,11]. In general, the purposes of modeling can be divided into three main groups: 1) elucidation of the mechanisms of interaction the elements of the system; 2) identification of model parameters from experimental data; 3) prediction of the behavior of the system under various external influences. In the framework of this paper we have interested in the accuracy of the data obtained, which can be attributed to the second group of modeling goals [12].

B. The biobject analysis

The object of research is the frontal part of the head bloodstream, including the ophthalmic arteries. Diagnostic data of blood pulse are obtained with the help of the developed electrodes system. The electrical impedance method is a technique for obtaining diagnostic information by means of non-invasive electrical sounding [13-15]. It is possible to visualize the internal structures of bioobjects using the electrical impedance method due to the fact that different tissues have different electrical conductivity. Blood flow especially affects the electrical conductivity of the tissue. This method has widely used to study the circulation of the brain [14].

The main problem arising in the development of electrode designs for electrical impedance measurement is the achievement of the required accuracy of the results [16-18]. To solve the problem, it is required to place the electrodes

around the area under investigation with an accuracy of 5 mm. The selection of the necessary positioning points and verification of the developed structures is carried out by computer simulation.

In this paper, we consider an 6-electrodes system, which is located at eye level (see Fig.1). As a result, it is possible to obtain data of blood filling of the internal carotid, middle cerebral, anterior cerebral and ophthalmic arteries. The diameter of the eye artery is the smallest of the recorded arteries (varies in the range from 0.3 to 0.5 mm) [19]. In this work we will carry out simulation for this artery.

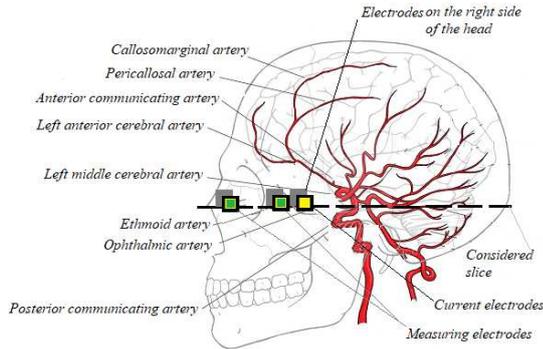


Fig.1. Scheme of electrodes position using the developed system

The ophthalmic artery lies outside of the optic nerve, moving on its way through the optic canal into the orbit [20]. The ophthalmic artery is located only in the front of the head, so we accept the first assumption - in modeling we will consider only the front of the head [21]. The second assumption is that the distribution of equipotential flow lines in the layers of tissues will be enough to obtain data on the pulse blood filling.

C. Selection of modeling tools

In electrical impedance methods, it is proposed to use high-frequency currents that will pass through the bioobject. Accordingly, the mathematical apparatus will describe electromagnetic interactions taking into account the boundary conditions [15].

Calculations will be made with the help of Netgen Mesh Generator. For calculation, it is necessary to determine electrical parameters of tissues. The main tissues considered in this simulation are bone, scalp, brain substance, eye, blood. Table I presents a summary of the set parameters [22].

TABLE I. RELATIVE PERMITTIVITY AND THE CONDUCTIVITY OF THE TISSUES

Layer name	Conductivity (S / m)	Relative permittivity
Scalp	0.5370	3300
Bone	0.0083	472
Brain substance	0.1340	3220
Blood	0.7030	5120
Eye	0.4990	1060

D. Construction of the geometry of the bioobject

The images obtained with the help of magnetic resonance tomography were analyzed to take into account all the layers entering into the mathematical model. The Fig.2 shows a snapshot reconstructed from MRI data (the NeuroImaging & Surgical Technologies Lab laboratory database) [23].

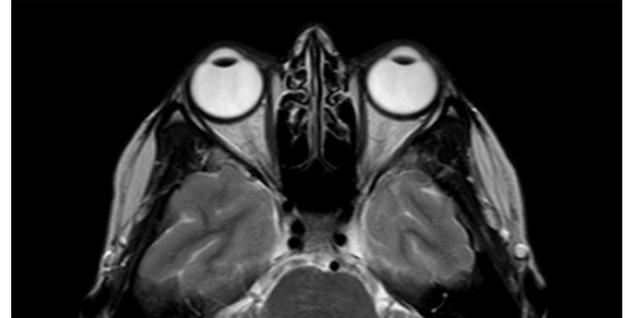


Fig.2. Snapshot of MRI with identification of the main biological structures of the eye

20 MRI snapshots from [23] were analyzed for the further reconstruction of the object geometry. The averaged geometry of the object was constructed on the basis of received data. The model is represented in the Fig.3. Since this model is intended only for determining the distribution of the probing current in the layers.

Six metal electrodes with a diameter of 5 mm are represented. The distance between the electrodes is fixed and has been chosen taking into account the anatomical structure of the skull (varies from 30 to 40 mm). The current source is located on the surface of the leftmost electrode; the current strength is 3 mA [24]. The electrodes located between them - the measuring electrodes (see Fig.3).

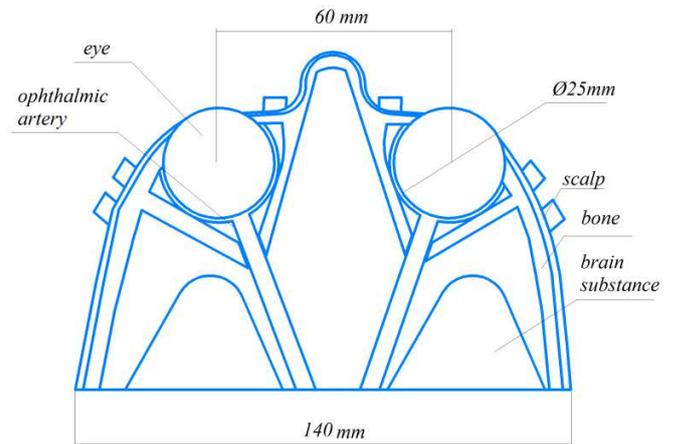


Fig.3. The geometry of the head front section in our modeling

E. Setting the basic simulation parameters

After downloading the obtained geometry of the bioobject into the software package, it is possible to specify parameters - the values of conductivity and relative permittivity for all layers.

The calculations have been carried out by the finite element method. The propagation of equipotential flow lines (see Fig.4) have been analyzed at time $t = 0$ s at an artery radius $r = 0.5$ cm.

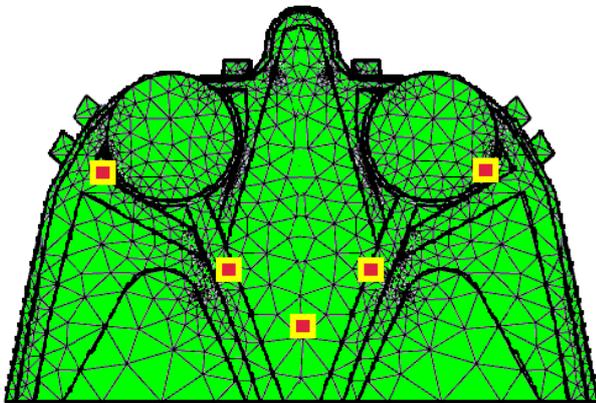


Fig.4. The distribution of equipotential lines in the our model

Due to the fact that the radius of the vessel varies with time due to the pulsations of the blood, we assigned its geometry as a variable value. To determine the required values, we used Doppler ultrasound signals. The Fig.5 shows a signal recorded from the ophthalmic artery by the Doppler ultrasound method (Signal Processing Laboratory database) [25].

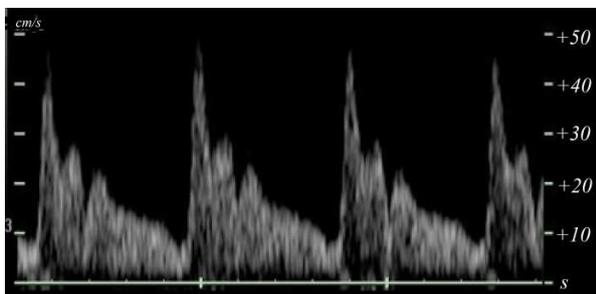


Fig.5. Doppler ultrasound signal of the ophthalmic artery

Doppler ultrasound shows a linear flow velocity in the vessel [25]. Pulsations of the radius of the blood vessel are associated with a change in the magnitude of the linear blood flow. Certain values of blood flow radius were stored in an array of values. The results obtained range from 0.35 to 0.50 mm per cardio interval. In our model, the geometry (radius) of the ophthalmic artery was changed within 1 second from 0.35 mm to 0.50 mm.

The Table II shows an example of the resulting array of current density distribution values for the example of 5 equipotential lines (considered points are indicated in the Fig.4 for one line).

TABLE II. CHANGE VALUE (IN A/CM^2) DEPENDING ON THE CHANGE OF THE VESSEL RADIUS

Line Number	1	2	3	4	5
1	0.44	0.37	0.30	0.37	0.44
2	0.41	0.38	0.29	0.38	0.41
3	0.45	0.37	0.31	0.37	0.45
4	0.44	0.38	0.29	0.38	0.44
5	0.46	0.37	0.30	0.37	0.46

Limit of permissible error of quantities is +/- 10%

RESULTS

The main parameter for studying the hemodynamics of the eye vessels is the calculation such a significant parameter as stroke volume. The values of the stroke volume obtained by the Doppler ultrasound method were compared with the results obtained on the basis of model values. The values of stroke volume were calculated for three cardio cycles. The Doppler ultrasound signals were taken from the database [25]. Table 3 shows the obtained values on one example.

TABLE III. COMPARISON OF STROKE VOLUME VALUES

Number of cardio cycle	Stroke volume (Doppler ultrasound), ml	Stroke volume (model volumes), ml
1	2.00	1.75
2	2.07	1.69
3	2.05	1.70

Limit of permissible error of quantities is +/- 10%

For a sample of 20 signals, the mean value of the RMS did not exceed 0.3, which may indicate the verification of the obtained model.

The practical application of the developed model can be the studying the relationship between blood filling of tissues and electrical conductivity. In particular, within the framework of this work, a study was made, based on the developed model, on the effect of changing the geometry of the object under consideration on electrical conductivity (Table 4).

TABLE IV. CHANGE IN RESISTANCE VALUE DEPENDING ON THE CHANGE OF THE VESSEL RADIUS

Pulse curve point	Radius of the vessel (mm)	Diameter of the eye (mm)	Obtained resistance value (Ohm)
The beginning of a cardiocycle	0.35	25.0	0.345
The systolic wave	0.50	25.3	0.178
The diastolic wave	0.40	25.1	0.227

In the second part of the study, the resistivity values were varied with the unchanged geometry of the object (Table 5).

TABLE V. CHANGE IN THE RESISTANCE VALUE DEPENDING ON CONDUCTIVITY

Pulse curve point	Conductivity of the eye, S/m	Obtained resistance value, Ohm
The beginning of a cardiocycle	0.4990	0.286
The systolic wave	0.5190	0.269
The diastolic wave	0.4890	0.263

Limit of permissible error of quantities is +/- 10%

Comparing the values of Table 4 and 5, it is obvious that the obtained values of resistances have significant differences. This may be due to fairly coarse changes in electrical conductivity. Thus, we can conclude that this model is workable, but it requires the specifications of certain parameters.

CONCLUSION

As a result of this work, a mathematical model has been developed to study the pulse blood filling of the anterior part of the head. In the future, it will allow to determine the levels of diagnostic signals, by changing the position of electrodes, to evaluate the accuracy of the obtained results.

REFERENCES

- [1] P. Neittaanmäki, S. Repin and T. Tuovinen (Eds.). *Mathematical Modeling and Optimization of Complex Structures*; Series: Computational Methods in Applied Sciences. Springer International Publishing AG, Switzerland; E-book, XXI, 2016.
- [2] C. Dym, *Principles of mathematical modeling*. Amsterdam: Elsevier Academic Press, 2004
- [3] S. Banerjee, *Mathematical Modeling: Models, Analysis and Applications*. N.-Y.: Chapman and Hall/CRC, 2014.
- [4] S. Pappalardo, "Mathematical modeling of biological systems", *Briefings in Bioinformatics*, Volume 14, Issue 4, 2013.
- [5] L. Formaggia, A. Quarteroni and A. Veneziani, *Cardiovascular mathematics*. Milan: Springer, 2009.
- [6] R. Brent, "A partnership between biology and engineering". *Nature Biotechnology*, vol. 22, no. 10, pp. 1211-1214, 2004.
- [7] P. V. Luzhnov, D. M. Shamaev, E. N. Iomdina, "Using quantitative parameters of ocular blood filling with transpalpebral rheoophthalmography". *IFMBE Proceedings* 65: pp.37-40, 2017.
- [8] P. V. Luzhnov, D. M. Shamaev, A. A. Kiseleva, E. N. Iomdina, "Analyzing rheoophthalmic signals in glaucoma by nonlinear dynamics methods". *IFMBE Proceedings* 68/2: pp.827-831, 2018.
- [9] O. Wolkenhauer, "The role of theory and modeling in medical research", *Frontiers in Physiology*, vol. 4, 2013.
- [10] M. Mark, *Mathematical modeling*. Edition: 3rd ed. Publisher: Singapore : Elsevier (Singapore), 2007.
- [11] T. Witeliski, M. Bowen. *Methods of mathematical modelling : continuous systems and differential equations*, Cham : Springer, 2015.
- [12] E. Tom, K. Schulman, "Mathematical models in decision analysis". *Infec Control Hosp Epidemiol*, vol. 18, pp. 65-73, 1997.
- [13] R Patterson, "Electrical Impedance Tomography: Methods, History, and Applications (Institute of Physics Medical Physics Series)", *Physics in Medicine and Biology*, vol. 50, no. 10, pp. 2427-2428, 2005.
- [14] Frerichs, J. Scholz and N. Weiler, "Electrical Impedance Tomography and its Perspectives in Intensive Care Medicine", *Intensive Care Medicine*, pp. 437-447, 2006.
- [15] A. Adler, R. Gaburro, W. Lionheart, "Electrical Impedance Tomography", in *Handbook of Mathematical Methods in Imaging*, 2nd ed O Scherzer (Ed), Springer, 2016.
- [16] D. M. Shamaev, P. V. Luzhnov, E. N. Iomdina, "Modeling of ocular and eyelid pulse blood filling in diagnosing using transpalpebral rheoophthalmography". *IFMBE Proceedings* 65: pp.1000-1003, 2017.
- [17] D. M. Shamaev, P. V. Luzhnov, E. N. Iomdina, "Mathematical modeling of ocular pulse blood filling in rheoophthalmography". *IFMBE Proceedings* 68/1: pp.495-498, 2018.
- [18] L. Callegaro, *Electrical impedance*. Boca Raton: CRC Press, Taylor & Francis Group, 2016.
- [19] R. D. Sinelnikov, Y. R. Sinelnikov, *Atlas of human anatomy: In 4 volumes. - 7 th ed., Rev. and additional. - T. 1. - Moscow: New Wave, 2007.*
- [20] Luzha D. *X-ray anatomy of the vascular system*. Budapest: Publishing House of the Hungarian Academy of Sciences, 1973.
- [21] D. M. Shamaev, P. V. Luzhnov, T. O. Pika, E. N. Iomdina, A. P. Kleyma, A. A. Sianosyan, "Applying transpalpebral rheoophthalmography to monitor effectiveness of the treatment of patients with glaucoma". *International Journal of Biomedicine* 6(4): pp.287-289, 2016.
- [22] C. Gabriel, *Compilation of the dielectric properties of body tissues at RF and microwave frequencies* / King.s College London. 1996.
- [23] "BITE: Brain Images of Tumors for Evaluation database – NIST", [Nist.mni.mcgill.ca](http://nist.mni.mcgill.ca), 2018. [Online]. Available: http://nist.mni.mcgill.ca/?page_id=672. [Accessed: 08- May- 2018].
- [24] C. Dimas, P. Tsampas, N. Ouzounoglou, and P. Sotiriadis, "Development of a modular 64-electrodes Electrical Impedance Tomography system", 2017 6th International Conference on Modern Circuits and Systems Technologies (MOCASST), 2017.
- [25] "Ultrasound image database | SPLab", [SPLab.cz](http://splab.cz), 2018. [Online]. Available: <http://splab.cz/en/download/databaze/ultrasound>. [Accessed: 08- May- 2018].

Unintended effects of dependencies in source code on the flexibility of IT in organizations

Debbie Tarenskeen
HAN University of Applied
Sciences, Arnhem, The
Netherlands
Email: debbie.tarenskeen@han.nl

Rogier van de Wetering
Open University, Heerlen, The
Netherlands
Email:
Rogier.vandeWetering@ou.nl

René Bakker
HAN University of Applied
Sciences, Arnhem, The
Netherlands
Email: Rene.Bakker@han.nl

Abstract—This study links business requirements and adaptability of existing software systems. Organizations expect flexibility of IT with regard to business requirements. We hypothesize that the flexibility of business requirements is difficult in IT systems, because of software dependencies in the way domain knowledge is implemented. In this paper, we, therefore, explore how Business requirements have been implemented in the source code of three open source healthcare systems. Outcomes suggest that a tight interdependency of business terminology and functionality in source code hides business requirements from view and thereby hinders IT flexibility on higher levels.

I. INTRODUCTION

SCHOLARS investigate strategic alignment of business and information technology (IT) for more than three decades within the information systems (IS) community. Recently, the importance of the role of flexible IT infrastructures for strategic alignment has been demonstrated anew for deployment, innovation, and evolution of IT systems in firms that operate in turbulent industries, including healthcare [1-4]. In the software architecture domain, software adaptability is seen as a quality attribute of software in general. Thereby meaning, for instance, the applicability of technological innovations or new technical features. Software adaptability is not explicitly aimed at changes in the business domain [5-7]. Expectations of the business do value general adaptability of systems, but also assume adaptability regarding business requirements. This current study focuses on changes in the business domain and business requirements and its consequences for software. We examine adaptability of IT systems regarding business terminology and business requirements.

II. IT FLEXIBILITY IN ENTERPRISE ARCHITECTURE

Although Enterprise architecture methods such as TOGAF focus on high-level business requirements, in the Business architecture, they are meant to show relations between high-level requirements and supporting architectures. We use definitions of TOGAF, because TOGAF aims at describing all levels of the IT infrastructure. TOGAF describes the supporting IT as data architecture, application architecture, and technology architecture. The definition of architecture in

this paper follows TOGAF's: "The fundamental organization of a system, embodied in its components, their relationships to each other and the environment, and the principles governing its design and evolution." Based on ISO/IEC 42010 according to TOGAF.

III. RELATED RESEARCH ON SOFTWARE EVOLUTION

Adaptability of software in empirical research can be positioned in the domain of research of Evolution of software. Within this domain, we notice that the evolution of business requirements is only marginally addressed.

Lehman strongly influences the research field of software evolution. The Laws of software evolution have been stated and evaluated during more than a decade of research [9, 10]. Research in this field has made no explicit distinction between the evolution of systems based on requirements in general, and evolution of systems based on new business requirements.

Numerous studies have emphasized the complexity of source code changes after the initial system has been realized, for example, see [11-13]. Studies that examine the relation of source code to IS architectures have a different focus than this research. They, e.g., aim at developing frameworks for software architecture evolution knowledge [14], or frameworks for classifying architecture-centric software evolution research [15], or on automatically updating architecture documents based on software changes [16].

IV. AXIOMATIC DESIGN AND CONCEPTUAL INDEPENDENCE

To present our point of view, we start by explaining the theoretical basis for adaptable and flexible low-level software components in an IT architecture. The theoretical views focus specifically on the adaptability of business requirements instead of on adaptability of software in general. The theoretical principles of Conceptual independence (CI), and the independence of functional requirements such as described in Axiomatic design (AD) [17-19] will be researched in this study in real-life software.

We report on a code mining study of open source code for Healthcare organizations for electronic health record systems

(EHR), to examine the way business terminology is applied. Then we will argue that there is a direct link of adaptability on the source code level to IT flexibility as expected by the Business architects.

V. RESEARCH QUESTIONS

We explore three selected open source software systems to find out if a separation of business terminology and application code has been effectuated, to create flexibility in the source code (CI). Hence, our first question concerns description of the software systems by the developers:

RQ1: Are indications of CI found in the documentation?

Then, we question the interdependency of the data model and the application source code. We implicitly assume that the data in the database model will represent the data that will be persistently stored.

RQ2: Does code demonstrate interdependency of table names and source code?

Next, we want to examine the flexibility of the specific healthcare terminology in the software system, based on CI. Thus, we define:

RQ3: Is CI applied in the software application?

Next, for AD functional requirements are primary. So we define:

RQ4: Does the source code of the system show different components that are related to separate Functional requirements?

The remainder of this paper is structured as follows. First, we highlight the theoretical aspects relevant to this study. Next, we present our methods section which is followed by the results section. We then discuss our findings and end with concluding remarks and some suggestions for future research.

VI. THEORETICAL BACKGROUND

A. Axiomatic design

The principles of AD are explained by Suh [19]. He explains how a design method should account for the independence of functional requirements and a low information density in different design parameters of the system. He calls these characteristics the Independence axiom and the Information axiom. The systems he describes are industrial systems, but he emphasizes that these principles can be applied to IT [20]. We interpret design parameters as design components of an IT system. The objective of AD is to realize systems that are flexible and understandable. With AD, the designs are iteratively developed and have the domain of customer needs, demands and requirements as a point of departure. From customer needs, functional requirements are inferred. In the design, the functional requirements are formulated independently from each other and can be changed in the design without affecting the rest of the design. The principle of Independence of functional requirements is at the fundament

of AD and can be compared to patterns in software engineering [21], such as separation of concerns. However, realizing independent functional requirements is not a priority in software engineering [22].

B. Conceptual independence

We advocate CI, the decoupling of healthcare terminology and domain models from software code to be able to alter healthcare terminology or domain models flexibly. We ground this choice on previous case studies that argue that AD principles are hard to implement in software systems, because of the interdependence of data models and the behavior of the system [23]. The interdependency of data models and application code has been extensively studied in IS research [24-27].

McGinnes points to the interdependence of data models and software application code as Conceptual dependence. He advises the decoupling of the data model and application functionality, meaning the behavior of the application [18]. McGinnes defines the Conceptual model as the structure of the information that is used in the business. Comparing the Conceptual model to the Domain model in UML, we find that in UML often behavior is added to the classes in the Domain, this is not the case in McGinnes' Conceptual model.

McGinnes adds behavior to Concepts by ordering Concepts in Archetypical categories, that applications can access. The applications have a responsibility to interpret the Archetypical categories. The applications add the behavior based on the specific Archetypical category. For instance, for "Location" the application knows that the instances of this category can be presented on a map.

C. Relation of Conceptual model to model-driven development

McGinnes describes the conceptual model as a (business) data structure that is used by the application. The meta-model, of the conceptual model, is fixed, the content is variable. These structures are comparable to MDD described by the OMG, Object management Group [28]. There are four levels of models, each function as a meta-level of the lower level. These levels are M0 to M3. McGinnes positions the conceptual model itself on level M1 as data.

We will address the meaning of these levels briefly.

M3 MOF	Defines a language for specifying a metamodel Example: MOF
M2 UML	Defines a language for specifying models Example: UML
M1 User Models	Defines a language that describe semantic domains Example: model of a problem domain
M0 Instance Models	Contains run-time instances of the model elements defined in a model

Fig. 1 Diagram of Modeling levels of OMG

The M1 level is the most important and most discussed modeling level in practice of software engineering. It shows categories or classes and the associations between them. The content consists of terms, for example, Person, Product, Order. The level is comparable with table names in RDBMSs. The model M0, on the lowest level, contains the instances of categories M1 stored. It is comparable to records in RDBMSs or instances of objects in programming languages. See Figure 1, published in a whitepaper explaining the different levels of Modeling of OMG [29].

The M2 level contains the description of model elements in a modeling approach; this is the meta-level of the description of, e.g., UML-models. The highest level (M3) contains a description of all (possible) modeling approaches. It is intended for comparing different modeling approaches [29].

The description of McGinnes of the model is at the M2 level, the model that concerns business concept types is an M1 level-model [30]. In this paper, we will not explore the similarities of OMG and McGinnes further. We think the challenge in system development lies in separating Conceptual models from behavior.

We argue based on ideas behind Conceptual dependence that CI is a prerequisite to being able to separate the different functional requirements from each other in the behavior part of the application [23].

D. Ampersand as illustration

We first, will describe a prototype system called Ampersand², based on a requirements specification language with relational semantics to illustrate the feasibility of implementing principles in source code [31]. Ampersand relies on model-driven development (MDD) to generate systems entirely defined by its domain model and business rules.

E. Ampersand applies CI and relation algebra for AD

We will explain the workings of Ampersand to demonstrate that flexibility of business functional requirements is feasible on source code level. This example is added for technical readers to explain the low-level code involved in separating business terms from application code. It also demonstrates with low-level code that a possibility to separate the business requirements from each other can be accomplished. The system Ampersand has separated the conceptual model from behavior. It, therefore, conforms to CI. It is based on relation algebra and has a mathematical structure [31, 32]. The conceptual model in Ampersand consists of concepts and relations between concepts. All information about concepts and relations is described in an Ampersand script (typically a .txt file). There is no extra

information of the business hidden in the software system. Behavior is described and defined in invariant (or declarative) business rules. The behavior is only applicable to the concepts and relations in the script. A script contains one Context that is entirely separate from other Contexts that can be defined in Ampersand. Ampersand applies rules as a way to connect the conceptual model structure to behavior. Rules can also be defined to check the consistency of data. Ampersand applies Rule checking behavior to define the software behavior. Rule checking is applied to Concepts and Relations in the Ampersand model. Examples from rules in healthcare can be: Diagnoses must have a relation to a Medical doctor, Diagnoses must have a date, a Patient cannot receive medication without the consent of the MD.

Each rule must be independent of the other rules in Ampersand, and therefore, behavior can be defined according to independent business functions.

F. Ampersand Runtime

The Ampersand Runtime can read, parse the script and import the Conceptual model, data, and rules. The script contains models on level M1 and M0. After reading this, a Rule engine checks business rules and signals violations. It operates on any script that conforms to the syntax and constraints of the Ampersand approach (On level M2).

G. Example Ampersand script

The following description of the Ampersand script is the model in natural language on level M2 of the OMG. Here we describe constraints and model elements (categories) that can be present in the script.

The first term in the Ampersand script is the word: CONTEXT. It signals the beginning of the script. ENDCONTEXT signals the ending. Then a PATTERN is presented consisting of CONCEPTS and RELATIONS.

After the pattern, the word ENDPATTERN closes this part, and in the script, PROCESSES can be defined regarding Ampersand RULES.

Summarizing, we can state that Ampersand follows the principles of CI by providing flexibility for the structure and naming of the data model. There are two different methods for keeping the conceptual model separate from the application code in Ampersand. First, the Ampersand Runtime works directly with the script and does not know about the domain in the script. Second, the script can be used for MDD. The Ampersand system conforms to the Independence axiom of AD, at least as far as functional requirements are concerned that can be defined in rules.

We have explained the workings of Ampersand in detail to demonstrate that flexibility of business functional requirements is feasible on source code level.

² Named after the ampersand symbol (&). According to Michels et al. the name refers to getting the best from both business and IT, i.e., achieving results from theory and practice alike, and realizing the desired results effectively and more efficiently than ever before.

VII. RESEARCH METHOD

A. Data collection procedure

We report the outcomes of code analysis of three systems. Two of these systems are frequently used in international health practice. The third system implements the standard of openEHR; a development we see more often these days. The latter claims to support different kinds of models for medical data. The research data are downloaded systems from GitHub. These were run locally to assess runtime dependencies and check if the source code is complete. Then, we have analyzed the documentation and the source code. We classified source code in types, the source code for libraries, the source code for initializing the database, source code for user interface frameworks and source code for business and other functionality in the software system. Only the last type of files have been examined in RQ2.

Since the idea of the paper is to evaluate open source systems in healthcare for application of CI and AD, we have searched for open source systems with an active community. The systems have been included in the references (websites and date) [33-36]. All of these are web applications that were run by us with an apache or tomcat server. Cabolabs is written in grails, openEMR in PHP and openMRS in java. All could operate with a MySQL database. The cabolabs openEHR download consists of 1351 files with 48 different extensions. The openEMR consisted of 12118 files with 130 extensions. The openMRS software has two downloads, the standalone consists of 723 files with 56 extensions. Because we also wanted to analyze the java code, we have also downloaded the core of openMRS with 1623 files with 40 different extensions.

B. A multistep approach

For each system, we applied a multistep approach including the following action: (1) running the systems locally, (2) analyzing all relevant and available documentation, (3) analyzing the directory structure, (4) extracting the data model from the MySQL database, (5) analyzing specific healthcare terms (see chapter VIII), (6) analyzing if database tables are hardcoded or generated for the particular system and (7) selecting of source code with (business) functionality (manually).

We incorporate specific methodological considerations and action (per research question) into the results sections.

VIII. SEPARATE ANALYSIS OF HEALTHCARE TERMS

We wanted to assess if application code depended upon hardcoded names in software code derived from healthcare. In the source code distinguishing between names that refer to healthcare terms and names that are necessary to follow the technical program flow, is difficult. Since we do not have expertise on healthcare terminology, but we wanted to signal the terms that were healthcare terms, we have asked independent reviewers and one healthcare professional

(psychiatrist) to review the names of database tables and list the names derived from or partly related to healthcare.

We have asked four reviewers that are not researchers or in any way related to the case studies. We have asked them to evaluate every table name in the three databases.

TABLE I.
EVALUATION OF HEALTHCARE RELATED TERMS IN TABLE

		openEHR	openEMR	openMRS
Number of tables		59	212	148
Number of tables with Health care related names	According to 3 of 4 reviewers	1	30	16
	According to Health care professional	2	62	50
Proportion of tables with Health care related names	According to 3 of 4 reviewers	0,02	0,14	0,11
	According to Health care professional	0,03	0,29	0,34

The scores of the healthcare professional have been registered separately also. For every open source program, there were two scores: the percentage of table names that three of the four reviewers labeled as healthcare related. Moreover, the separate score of the healthcare professional.

In Table I we find the number and percentage of table names with recognizable healthcare terminology parts. The healthcare professional classified more names as healthcare related than the other persons, but all the table names that the 3 out of 4 persons listed were a subset of the names that the healthcare professional listed. We are now able to assess interconnectedness of application code to hardcoded healthcare terms.

IX. RESULTS CONCEPTUAL INDEPENDENCE AND AXIOMATIC DESIGN IN SOURCE CODE

A. Conceptual independence in the documentation

This section addresses RQ1. We have extensively read the associated documentation and searched for indications that the system is adaptable based on healthcare terminology. Through our analyses, and also based on our review of the Information model of openEHR (on <http://www.openehr.org/>), we conclude that openEHR indicates CI. A quote from documentation of openEHR confirms this view [37]: “Your EHR system does not need to know a priori about any of the clinical data it will process, such as vital signs, diagnoses or orders. Models for those things are developed separately. Models for data sets and forms are also developed separately, and UI form components are generated from these definitions.”

The data structure is said to be very flexible and can support transformations to other healthcare terminology standards.

In openEMR, there exists no reference to a model, but we find a description of the Database structure [38]. There is some variability for the conceptual model, by which we mean, the user can define categories in one table. Thus, openEMR is partly flexible concerning terminology, but not concerning models of healthcare data. Finally, openMRS shows signs of CI and AD. To highlight this particular view, we quote from the wiki documentation of openMRS [39]: *“At the heart of OpenMRS is a concept dictionary. This dictionary, much like a typical dictionary, defines all of the unique concepts (both questions and answers) used throughout the system.”*

The software itself, openMRS, in essence, is constructed to support ‘modules.’ Implementations can modify the behavior of the system to meet local requirements using these modules. Because changes can be added to the Conceptual model, it is not necessary for everyone having to agree on a single approach.

B. Interdependency of table names and source code

We now report the results of for RQ2. These results consist of totals of code mining results.

We have defined two indications of the interdependency of code and data, i.e., I) hardcoded use of table names in the source code of more than 80 percent of the tables and II) hardcoded use of table names in the source code of more than 80 percent of the source code.

We did not find the expected first indication in every system. We would have expected the use of all the tables in the source code. If table names are missing, they are not used. Since the named applications are the only applications that use the database tables, this needs further investigation.

The second indication has been signaled in the source code files. If names of database tables in source files are hardcoded, then any particular change in table names implies changes in the source code. With table names spread over different files, then changes in table names lead to changes in multiple maybe interdependent files, leading to unpredictable behavior of the software application.

Why would a table name change? If ideas about the Conceptual model change or if other functionality needs other data concepts or maybe extra attributes (columns in tables) then the table names and columns will change. Adaptations of concepts and table names or columns are frequent in the evolution of code [25]. In this empirical study of Qiu, it was found that adding tables, columns and changing names of tables and columns frequently appear. We focus here, specifically, on the names that are healthcare related, because we signal a relation between business terms and source code files.

Our method can be described in the following way: we have extracted all table names from the applied database management system and have counted the number of times these names are hardcoded in software source code. We have calculated the percentage of database table names that were found in the source code files. We have counted the number of source code files that access database table names directly, or use Class names that are derived from table names, for instance by removing the dash. In Figure 2 the relations of the source code files to table names are visually presented. We also have calculated the percentage of source code files, that access table names directly.

Counting will demonstrate the relation.

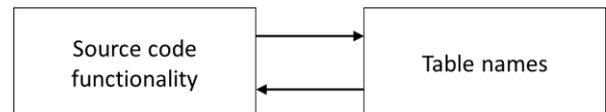


Fig. 2 The existence of relations between source code and table names

Concerning openEHR, 46% of table names have been found in the programming code, but only two of those are marked as Healthcare related. The names are doctor-proxy and patient-proxy, but no table names are related to medical knowledge. The groovy files with table names accounted for 99% of 176 files, but these were not marked as Healthcare related, exception above. In groovy files, 69% class names have been found, that are derived from table names. Groovy files with these class names accounted for 69% of the groovy files.

In the openEMR download, 82% of table names are found hardcoded in PHP-code. Including 29 of 30 with Healthcare-related table names. In the PHP-files 94% of 5401 files have access to hard-coded table names.

In the openMRS-core download, we have found 57% of table names are hardcoded in java-source code. Including almost all table names (14 of the 16), that have been marked Healthcare related. In the Java-files in the openMRS-core, 100% of 1019 files access hardcoded table names.

Based on the second indication, we find an extensive interdependency between source code and database table names in all three systems.

X. IS CONCEPTUAL INDEPENDENCE APPLIED IN THE SOFTWARE APPLICATION?

This paragraph reports results for RQ3. The indications below are derived from characteristics of CI:

- *Indication: No hardcoded use of healthcare-related table names in the source code.*
- *Indication: A presence of a separate structured model for healthcare terms, in the source code for generating database tables.*
- *Indication: A presence of a separate structured model for healthcare terms, in a separate file.*

We expect that when the healthcare-related table names are found in source code files, then changes in healthcare terminology directly affect the source code.

So for the first indication of RQ3, we have to mark database table names that have a direct reference to the medical terminology in the system. For this indication, we had first to classify all table names in “Unknown name” and “Healthcare terminology name.” See Chapter VIII. Then we searched for occurrences of healthcare-related table names in source code files. Two systems: openEMR and openMRS, applied hardcoded healthcare-related table names in the source code.

The other two indications, above, are meant to demonstrate a separation of the conceptual model and the application code, as is a characteristic of CI. In detail, we have found that openEMR and openMRS have applied frameworks for separating business domain terms (the Conceptual model) and business logic from the rest of the application code. The frameworks used are Zend for openEMR and Hibernate for java in openMRS. These frameworks and the related source code of the systems have been analyzed. The frameworks use script code for defining the (Business) Conceptual model. They do not separate the Conceptual model from the behavior of the software. Therefore these do not comply with CI. The frameworks aid the developers with building and partly generating source code. Hibernate helps developers in separating database management systems from source code but does not aid in decoupling business terminology from source code. The framework script code then becomes part of the source code. We cannot directly extract the applied Conceptual models.

The software of openEHR contains separable Conceptual models apart from application logic. We confirm the existence of separate Conceptual models because we also find “parsers” and “indexers” in the source code.

For the last indication, we have counted the number of times table names can be found in one file, to search for indications of a definition file for the Conceptual model. In openEHR, we found the “opt-file” and “adl-file,” in which the M1 model is included as data. They comply with the M2 model of openEHR. Therefore it can be used for separating the Conceptual model from the behavior of the software.

In the openMRS source, the liquibase tool is applied for updating tables based on changes in the database for new modules. With the liquibase functionality updates on database structure and data can be automated with liquibase.xml-files. The M1 model is input as xml-data, but no M2 model can be found.

In openEMR, only an .sql file was found that contained 188 of the 212 tables. The healthcare related terms are not included as data but are hardcoded in sql. It cannot be used as an M1 model, because changing it will break the source code and no M2 model can be found.

Concluding: In the source code of Cabolabs openEHR server system all three indications have been found. Several files with the Conceptual model and its instances (M1 and M0) have been signaled. These files with extensions .adl and .opt can be reused by other openEHR standard based systems. Cabolabs openEHR-server applies CI.

In the other systems frameworks such as Hibernate and Zend have been used, for partly separating the model from the application code. Further, the frameworks do not distinguish business terms from application code classes. The consequence is that the current application of frameworks involves code programmers for adaptation of business logic and business terminology.

XI. AXIOMATIC DESIGN APPLIED IN SOURCE CODE

In this paragraph, a report of RQ4 is given. For AD, functional requirements are primary. In AD it is required that the software system can be divided into components that are related to functional requirements. Systems based on AD will be adaptable based on changing functional requirements because business IT architects can pinpoint specific source files where changes are necessary.

RQ4 will lead to demarcation lines in the Runtime components or demarcation lines in the source code, which has different independent functional requirements.

- *Indication: Existence of directory structures in the source code that show Functional requirements*
- *Indication: Existence of runtime modules that can be added and deleted for Functional requirements behavior that is executed*

The indications for Axiomatic design will be studied in detail in future research. In this overall check of the source code, it is found that openMRS contains a directory structure for separate modules. We find complementary functionality in the openMRS runtime application because modules with Business functionality can be turned off and on. With the openEHR server software, tooling is under construction that can generate User interfaces based on the opt-files. Moreover, thus separation of high-level functional requirements can be realized.

XII. CONCLUSION AND DISCUSSIONS

In this study, we have explored the adaptability of source code concerning the business requirements and changes in the business domain terminology. Interdependency of the data model and the application code can make systems hard to change, this is seen in the literature and in our investigation of open source healthcare systems.

However, the dependency of the application source code on healthcare terms can be avoided by separating these terms in a separate model as input for the application. We demonstrate this with the Ampersand prototype, where indications for CI and AD can be located in the source code.

We have explored how business terms and business functionality appear in application source code in three open source systems actively used in healthcare. We have shown that CI, separating the business terms from the application software, can be applied and is applied in openEHR and partly in openMRS. AD has not been studied extensively in this case, but indications for AD are found in openMRS.

We conclude that because of the extensive interdependence of the data model and application source code in openEMR and openMRS, business terminology becomes part of the source code and cannot be adapted without radically changing the source code. So we conclude that in these systems the flexibility of business terminology is obstructed if the business terminology is not explicitly separated from the application source code.

Despite this studies contributions, there are several limitations that future research should address. The researchers remark that an alternative to separating the conceptual model from application source code would be to use tooling for source code editing based on business requirements. Moreover, currently, some indications for this kind of tools were present in the source code that is examined. Frameworks help to separate the Conceptual model from the application code, but in the end, Conceptual models become an integrated part of the source code. When the Conceptual model is included in the source code, then it will depend on professional skills or discipline of the programmer(s), to check that the Conceptual model will stay separated from application code. Since frameworks do not distinguish health care terms from software application classes, medical expertise is necessary to locate these.

In this paper, we have only studied software architecture for a limited number of applications and components. Therefore it can be questioned if a full-scale application of this principle can be implemented in enterprise architectures. We are currently researching the design and implementation of a separate (conceptual model-layer) data layer in a large scale healthcare IT architecture.

REFERENCES

- [1] J. C. Henderson and H. Venkatraman "Strategic alignment: Leveraging information technology for transforming organizations," *Ibm Systems Journal*, 1, (1993), 32, pp. 472-484, doi: 10.1147/sj.382.0472
- [2] R. Van de Wetering, P. Mikalef and A. Pateli "A strategic alignment model for IT flexibility and dynamic capabilities: toward an assessment tool," *Proceedings of the 25th European Conference on Information Systems (ECIS)*, Guimarães, Portugal(2017), pp. 1468-1485, doi:
- [3] R. van de Wetering, P. Mikalef and R. Helms "Driving organizational sustainability-oriented innovation capabilities: a complex adaptive systems perspective," *Current Opinion in Environmental Sustainability*(2017), 28, pp. 71-79, doi: <http://dx.doi.org/10.1016/j.cosust.2017.08.006>.
- [4] R. van de Wetering, J. Versendaal and P. Walraven "Examining the relationship between a hospital's IT infrastructure capability and digital capabilities: a resource-based perspective, doi:
- [5] L. Bass, P. Clements and R. Kazman *Software architecture in practice*. Addison-Wesley Professional, Upper Saddle River, NJ, 2012.
- [6] F. Bachmann, L. Bass, D. Garlan, J. Ivers, R. Little, P. Merson, R. Nord and J. Stafford *Documenting Software Architectures: Views and Beyond*. Addison-Wesley Professional, 2011.
- [7] J. Tyree and A. Akerman "Architecture decisions: Demystifying architecture," *Ieee Software*, 2, (2005), 22, pp. 19-+, doi: 10.1109/ms.2005.27.
- [8] TheOpenGroup *TOGAF Version 9.1 Evaluation copy*. The Open Group, 2011.
- [9] M. M. Lehman "Laws of software evolution revisited," *European Workshop on Software Process Technology*(1996), pp. 108-124, doi: <https://doi.org/10.1007/BFb0017737>.
- [10] I. Herraiz, D. Rodriguez, G. Robles and J. M. Gonzalez-Barahona "The evolution of the laws of software evolution: A discussion based on a systematic literature review," *ACM Computing Surveys (CSUR)*, 2, (2013), 46, pp. 28, doi: 10.1145/2543581.2543595.
- [11] N. Ajiienka, A. Capiluppi and S. Counsell. "Managing Hidden Dependencies in OO Software: a Study Based on Open Source Projects." In *Proceedings of the Empirical Software Engineering and Measurement (ESEM)*, 2017 *ACM/IEEE International Symposium on*, IEEE, 2017, pp. 141-150, doi: 10.1109/ESEM.2017.21.
- [12] H. Kagdi, M. L. Collard and J. I. Maletic "A survey and taxonomy of approaches for mining software repositories in the context of software evolution," *Journal of Software: Evolution and Process*, 2, (2007), 19, pp. 77-131, doi: 10.1002/smr.344.
- [13] H. Kagdi and D. Poshyvanyk "Who can help me with this change request?," *Program Comprehension*, 2009. *ICPC'09. IEEE 17th International Conference on*(2009), pp. 273-277, doi: 10.1109/ICPC.2009.5090056.
- [14] A. Ahmad, P. Jamshidi and C. Pahl "A framework for acquisition and application of software architecture evolution knowledge: 14," *ACM SIGSOFT Software Engineering Notes*, 5, (2013), 38, pp. 1-7, doi: 10.1145/2507288.2507301.
- [15] P. Jamshidi, M. Ghafari, A. Ahmad and C. Pahl. "A framework for classifying and comparing architecture-centric software evolution research." In *Proceedings of the Software Maintenance and Reengineering (CSMR)*, 2013 *17th European Conference on*, IEEE, 2013, pp. 305-314, doi:
- [16] T. Haitzer, E. Navarro and U. Zdun "Reconciling software architecture and source code in support of software evolution," *J Syst Softw*(2017), 123, pp. 119-144, doi: <https://doi.org/10.1016/j.jss.2016.10.012>.
- [17] S. McGinnes. "The Problem of Conceptual Incompatibility." In *Proceedings of the International Conference on Availability, Reliability, and Security*, Springer, 2011, pp. 69-81, doi:
- [18] S. McGinnes and E. Kapros "Conceptual independence: A design principle for the construction of adaptive information systems," *Information Systems*(2015), 47, pp. 33-50, doi: <https://doi.org/10.1016/j.is.2014.06.001>.
- [19] N. P. Suh *Axiomatic Design: Advances and Applications (The Oxford Series on Advanced Manufacturing)*. Oxford University Press, New York Oxford, 2001.
- [20] N. P. Suh "Fundamentals of Design and Deployment of Large Complex Systems: OLEV, MH, and Mixalloy," *Journal of Integrated Design & Process Science*, 3, (2012), 16, pp. 7-28, doi: 10.3233/jid-2012-0001.
- [21] C. Larman *Applying UML and patterns : an introduction to object-oriented analysis and design and iterative development*. Prentice Hall PTR Upper Saddle River, N.J., 2005.
- [22] F. Buschmann, K. Henney and D. Schmidt *Pattern-oriented Software Architecture: on patterns and pattern language*. John wiley & sons, 2007.
- [23] D. Tarenskeen and R. Bakker. "Applying Axiomatic design and Conceptual independence in the domain of IT systems." In *Proceedings of the ICAD 2017 International Conference on Axiomatic Design*, Iasi Romania, 2017, doi: <https://doi.org/10.1051/mateconf/201712701006>.
- [24] A. Aryani, F. Perin, M. Lungu, A. N. Mahmood and O. Nierstrasz. "Can we predict dependencies using domain information?." In *Proceedings of the Reverse Engineering (WCRE)*, 2011 *18th Working*

- Conference on, IEEE, 2011, pp. 55-64, doi: <http://doi.ieeecomputersociety.org/10.1109/WCRE.2011.17>.
- [25] D. Qiu, B. Li and Z. Su. "An empirical analysis of the co-evolution of schema and code in database applications." In Proceedings of the Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, ACM, 2013, pp. 125-135, doi: 10.1145/2491411.2491431.
- [26] A. Cleve, M. Gobert, L. Meurice, J. Maes and J. Weber "Understanding database schema evolution: A case study," Sci. Comput. Program.(2015), 97, pp. 113-121, doi: <https://doi.org/10.1016/j.scico.2013.11.025>.
- [27] T. Mens, L. Meurice, M. Goeminne, C. Nagy, A. Decan and A. Cleve "Analyzing the Evolution of Database Usage in Data-Intensive Software Systems, October 14, 2017, (2017), 2017, doi: [28]
- [28] I. Object Management Group Meta Object Facility™ (MOFTM) Core 2.5.1. 2016. <http://www.omg.org/spec/MOF/>. Retrieved October 9, 2017, Accessed in 2017.
- [29] I. Object Management Group Meta-Modeling and the OMG Meta Object Facility (MOF) . 2017. www.omg.org/ocup-2/documents/Meta-ModelingAndtheMOF.pdf. Retrieved October 9, 2017, Accessed in 2017.
- [30] J. Bézivin and O. Gerbé. "Towards a precise definition of the OMG/MDA framework." In Proceedings of the Automated Software Engineering. (ASE 2001). 16th Annual International IEEE, 2001, pp. 273-280, doi: 10.1109/ASE.2001.989813.
- [31] G. Michels, S. Joosten, J. van der Woude and S. Joosten. "Ampersand." In Proceedings of the International Conference on Relational and Algebraic Methods in Computer Science, Springer Verlag, 2011, pp. 280-293, doi: DOI https://doi.org/10.1007/978-3-642-21070-9_21.
- [32] G. Michels, S. Joosten, J. v. d. Woude and S. Joosten. "Ampersand applying relation algebra in practice." In Proceedings of the Proceedings of the 12th international conference on Relational and algebraic methods in computer science, Rotterdam, The Netherlands, Springer-Verlag, 2011, pp. 280-293
- [33] P. Pazos openEHR cabolabs server-v0.9. 2017. <https://github.com/ppazos/cabolabs-ehrserver>. Retrieved 03/05/2017, Accessed in 2017.
- [34] openEMR openEMR-v5.0.0. 2017. <https://github.com/openmrs/openmrs-standalone>. Retrieved 03/03/2017, Accessed in 2017.
- [35] openMRS openMRS Core-v4.0.0. 2017. <https://github.com/openmrs/openmrs-core>. Retrieved 03/04/2017, Accessed in 2017.
- [36] openMRS openMRS Standalone 2.5. 2017. <https://github.com/openmrs/openmrs-standalone>. Retrieved 03/03/2017, Accessed in 2017.
- [37] What is openEHR? 2017. http://www.openehr.org/what_is_openehr#. Retrieved October 14, 2017, Accessed in 2017.
- [38] openEMR Database structure openEMR. 2014. http://www.openemr.org/wiki/index.php/Database_Structure. Retrieved October 14, Accessed in 2017.
- [39] B. Mamlin and S. Jindal Introduction to OpenMRS. 2017. <https://wiki.openmrs.org/display/docs/Introduction+to+OpenMRS>. Retrieved October 14, 2017, Accessed in 2017.

ECG signal coding methods in digital systems

Tomasz Żentara
Military University of Technology
Urbanowicza Str. 2,
00-908 Warsaw, Poland,
email: tomasz.zentara@wat.edu.pl

Krzysztof Murawski
Military University of Technology
Urbanowicza Str. 2,
00-908 Warsaw, Poland
IEEE Member # 92707852
email: krzysztof.murawski@wat.edu.pl

Abstract - Article contains an overview of ECG signal coding methods. The presented methods are used to record and present the raw ECG signal in digital systems. The aim of the presentation is to choose the best technique for use in the ECG recording device, currently being developed by the authors.

I. INTRODUCTION

The study of electrical activity of the heart using an electrocardiograph (ECG) has become one of the basic research used in medicine in the diagnosis of many heart diseases over time [1,2]. The first ECG recorders used the needle's movements ended with the marker, which marked on the moving tape the electrical indications of the heart. Nowadays, the development of digital systems, as well as the miniaturization of computer systems and microcontrollers allowed to significantly increase the possibilities, while reducing the size and weight of ECGs devices. At the same time, a "revolution" took place, consisting in replacing the paper record for portable memory as well as adding, often even built-in, displays or screen monitors.

The integration of medical and contextual data obtained from the patient into the structure of a single source file or message is very important in case of operating of digital systems used to process medical signals. At present, a number of companies producing ECG monitoring devices have their own methods of coding the signal of cardiac activity [3]. It is worth noticing that standardization organizations have already presented defined file formats and protocols describing medical information [4]. They usually contain elements describing the patient, e.g.: code, name, surname, weight, height, etc., and ECG biosignals or data from the encephalograph (EEG). General characteristics and activities of individual protocols are widely analyzed [5-7]. One of many examples is the SCP-ECG protocol that integrates ECG, EEG and carbon dioxide (CO₂) sensor signals. [6]

II. MOTIVATION

The article presents and characterizes the standards used in coding biomedical signals - mainly due to the authors' interest in the ECG signal.

The motivation for the work was to present and describe the currently used ECG signal coding standards. What is an important step to develop a prototype of own device used to measure ECG signal, in which selected standards will be implemented and tested.

III. LITERATURE RESEARCH

During the preparation of the publication, 16 literature items were analyzed, including 2 books [1,2], 6 articles and publications in scientific journals [3,5-7,11,12], as well as five generally accessible instruction manuals and user instructions [4,13-16].

As a result of literature analysis, the most important standards were specified, including SCP-ECG, MFER, HL7 and DICOM. They have been analyzed in terms of the implementation possibilities in the constructed ECG device. The conclusions from the analysis are included in this publication.

IV. MEDICAL SIGNALS PRESENTATION IN DIGITAL SYSTEMS

The basic problem identified in the field of bioengineering lies in a different look at the medical device by engineering and medical personnel. Doctors and medical staff are mainly focused on the visualization of the ECG signal. An example of such signal and its specific characteristic fragments as well as time intervals between them is shown in Fig. 1. It is analyzed by doctors who, as a result of its analysis, can determine a number of heart diseases or irregularities of its work.

The article focuses on the engineering area. Therefore, the form and accuracy of the ECG signal recording was more important than the analysis of its shape. The main reason for

This work was not supported by any organization

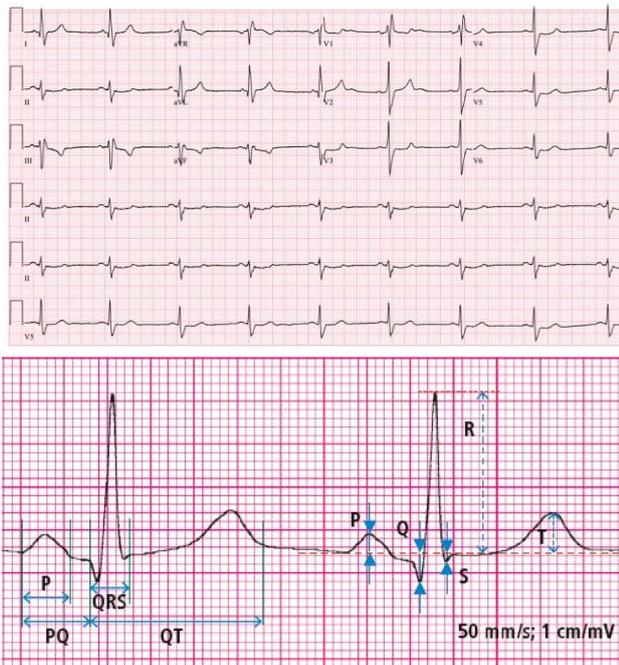


Fig. 1. ECG signal recording along with key elements and their markings [8,9]

this interest is the fact, that sensors connected to the patient's body register only an analog electrical signal. The signal presented in this form can not be saved in the memory of the digital device. This problem is solved on the way of conversion to the digital form. The analog / digital converters are responsible for signal conversion. Usually, they are fast 24 bit Δ / δ transducers that ensure signal conversion and sample representation accuracy at single bit level. This means in practice that the processing speed is so large that each of the two neighboring samples can differ only at the level of one bit. The signal prepared in this way is stored in the digital device and subjected to standardization and / or normalization.

V. CONCEPT FOR THE PROTOTYPE DEVICE

The project of the prototype is based on the use of the Texas Instruments ADS1298 ECG SOC specialized Front-End system. The ECG measuring electrodes will be connected to the connector, after which the signal will go through the set of filters and then to the ADS1298 system. This system will convert the analog signal into a digital form so that the evaluation set, equipped with the ARM processor, can read the raw ECG data. Next, for the data presentation the LCD is planned to use, as well as the recording of these data on an removable flash memory. In addition, these data will be sent via a radio link directly to the PC. Hence, so much emphasis on the choice of the appropriate coding standard. The concept device diagram is shown in Fig. 2.

Due to the newly-introduced regulation enforced in Poland related with personal data protection (polish shortcut RODO), for research purposes, it is planned to use mainly

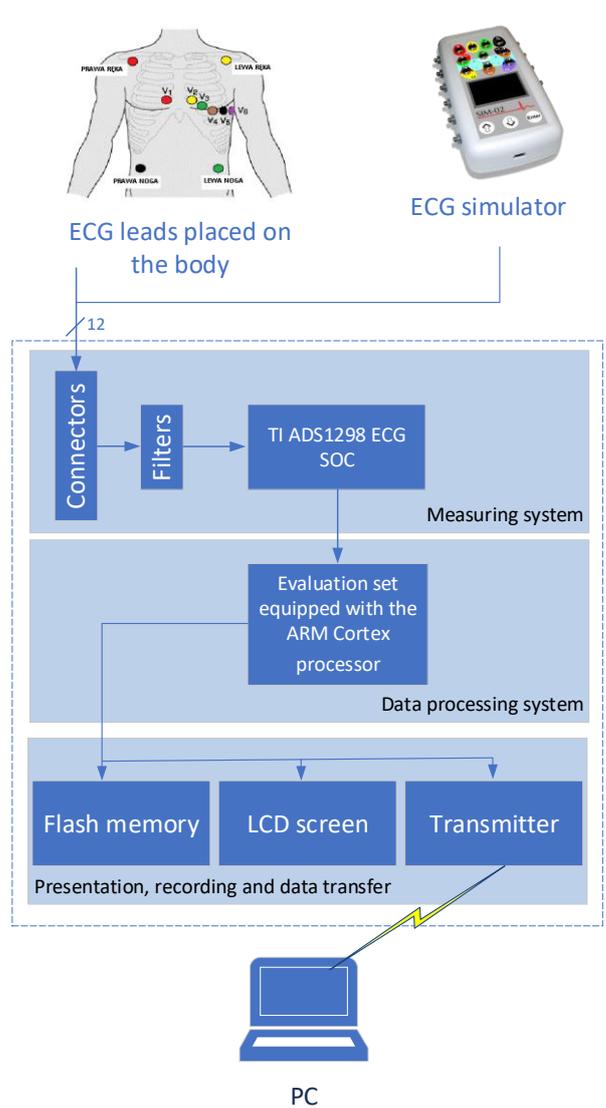


Fig. 2. Concept for the prototype device

artificial patient (ECG simulator), which will be generating the test data, as well as anonymous volunteers.

VI. DESCRIPTION OF THE SPECIFIC STANDARDS

The first of the described ECG signal coding standards is the *Standard Communication Protocol for Electrocardiography (SCP-ECG)*. It was developed as part of the implementation of the European ECG project in 1989-1991. In 1993, it was approved as the ENV 1064 standard. Sources report [5] that this standard is now a non-expanding standard, however other sources [6,7] question this statement by saying that it has been revitalized and continues to function only in the new version.

In both cases, the SCP-ECG standard imposes the requirement to use the structural form of cardiological information. This form adapts to the specific pattern of the usually imposed sections. A list of sections for the older SCP-ECG standard is shown in Table 1. The form of the newer standard is shown in Table 2.

Comparison of Tab. 1 and Tab. 2 shows the difference between the newer (Table 2) and the older (Table 1) version of the SCP-ECG standard. In the newer version, seven optional sections have been added. However, section 4 corresponding to the old standard for the location of QRS complexes (allowing for further subtraction of the reference rhythm to calculate the "residual signal") was reserved in the new version. In the older version of the SCP-ECG standard (Table 1), the section description is presented as follows. As the first one there is a section of indicators, which is a specific list of contents for the entire SCP record and can be used to identify the contents of the record. The next section, containing header information, can consist of up to 35 tags describing a number of patient data, including: identification number, first name, last name, age, weight, height, etc., as well as information on the medicines to be taken, reasons for referral, the apparatus used to measure, along with the software version of these apparatus. In addition, in the header section it is possible to write a text history about the subject of the patient's illness.

The next (in this case optional) section 2 stores information on how to encode the ECG signal, which is recorded in the following sections (5 and 6). In section 2

Tab. 1. The record structure for the older version of the SCP-ECG standard [5]

Name	Field status	Content
	Mandatory	Checksum (CRC) of the entire record (except this field) - 2 bytes
	Mandatory	The entire record size is 4 bytes (unsigned)
Section 0	Mandatory	Pointers to data areas in the record
Section 1	Mandatory	Header information - patient data/ECG - Acquisition data
Section 2	Optional	Huffman tables used in encoding of ECG data (if used)
Section 3	Optional	ECG leads definition
Section 4	Optional	QRS location (if reference beats are encoded)
Section 5	Optional	Encoded reference beat data if reference beats are stored
Section 6	Optional	„Residual signal” after reference beat subtraction if reference beats are stored, otherwise encoded rhythm data
Section 7	Optional	Global ECG measurements
Section 8	Optional	Textual diagnosis from the "interpretive" device
Section 9	Optional	Manufacturer specific diagnostic and overreading data from the "interpretive" device
Section 10	Optional	Per lead ECG measurements results
Section 11	Optional	Universal statement codes resulting from the interpretation

there is also a recoding table for lossless Huffman compression. Another, this time the third section contains a definition of the number of leads used to measure the electrical activity of the heart. The SCP-ECG standard permits the use of up to 255 measuring electrodes that can be distributed over the entire patient's body surface. However, the most common sets are 12 leads for ECG electrodes. For the most commonly used electrode sets, preliminary code lists of lead systems have been prepared as standard. It is also worth mentioning that section 3 contains information about the length and simultaneity of records for individual leads, so that the raw signal recording from individual electrodes can have different lengths and does not have to be made at the same time. Section 4 is responsible for storing position indicators for representative QRS wave units recorded in section 5. Section 6 in the SCP-ECG standard is intended for storing the signal after subtracting the representative units recorded in section 5 or the raw heart rhythm electric signal read from the cardiograph electrodes. The filling of section 7 requires more advanced devices that allow detection of the remaining electrocardiogram waves. For storing data for section 8, it is required to use devices equipped with diagnostic algorithms to classify the evolution of the heart. Sections 7 to 11 are classified for implementation in advanced electrocardiographs equipped with algorithms for recording interpretation, and also in efficient computer components.

Compared with the older version of the SCP-ECG standard (Table 1), the main changes in the newer version (Table 2) are the disappearance of the bimodal compression scheme and the subtraction of representative units that apply to sections 4 to 6. Lossless compression (Huffman coding) of short-term ECG data rhythms (section 6) and representative 'type 0' units (section 5) is allowed. Additionally, to simplify decoding, it is recommended in the new standard to save all ECG signal data as uncompressed, in the form of a series of integer lengths with a sign. However, in exceptional situations, e.g.: slow network data transmission or limited wireless connectivity, for mobile devices the standard continues to use Huffman coding and the differentiation of representative signals.

As mentioned earlier, in the SCP-ECG version 3.0 standard, new sections have been added to the record, and so the standard now provides the option of storing long-term ECG signal (section 12), e.g. for 3-lead ECG at 200 samples / s with a 16-bit resolution, continuous recording up to 40 days is possible. The format of section 12 is very similar to the ISHNE [11] format used to record data in Holter devices. Additional sections 13 and 14 have been introduced to support stress test records and drug trials, where section 13 contains metadata, or representative assemblies (or pointers for selected reference units), and section 14 contains several selected medium or short-term ECG sequences. In section 15 it is possible to store several predefined global measurements and lead measurements together with annotations to them.

Tab. 2. The record structure for the SCP-ECG standard in the newer version 3 [10]

Field status	SCP-ECG record structure
Mandatory	Checksum (CRC) of the entire record (except this field) - 2 bytes
Mandatory	The entire record size is 4 bytes (unsigned)
Mandatory	(Section 0) Pointers to data areas in the record
Mandatory	(Section 1) Header information - patient data/ECG Acquisition data
Optional	(Section 2) Huffman tables used in encoding of ECG data (if used)
Mandatory	(Section 3) ECG leads definition
Mandatory	(Section 4) Reserved for legacy SCP-ECG versions
Optional	(Section 5) Encoded type 0 reference beat data (if reference beat is stored)
Optional	(Section 6) Short-term ECG rhythm data
Optional	(Section 7) Global ECG measurements
Optional	(Section 8) Textual diagnosis from the "interpretive" device
Optional	(Section 9) Manufacturer specific diagnostic and overreading data from the "interpretive" device
Optional	(Section 10) Per lead ECG measurements
Optional	(Section 11) Universal statement codes resulting from the interpretation
Optional	(Section 12) Long-term ECG rhythm data
Optional	(Section 13) Stress tests, drug trials and protocol-based ECG recordings metadata
Depends on section 13	(Section 14) Selected ECG sequences repository
Optional	(Section 15) Beat-by-beat ECG measurements and annotations
Optional	(Section 16) Selected ECG beats measurements and annotations
Optional	(Section 17) Spikes measurements and annotations
Optional	(Section 18) Additional ECG annotations

The recorded sequences may refer to single beat selected by a doctor or a sequence typing algorithm (depending on the progress of the device / selected scenario), or to the whole set of beats from one / many time windows recorded in long-term section 12, short-term section 6 or data from section 14. The next element is section 16, which provides solutions for

storing another set of measurements and annotations than those recorded in the previous section. The structure and format of section 16 are the same as in section 15, so you can say that both sections are complementary with each other. However, in section 16 there is no restriction on determining the time window. Section 16 is preferred for storing heart beat measurements if "beat by beat" measurements are not required (if section 15 is not present). Another optional section, with number 17, has been designed to provide support for predefining and storing large global and / or pre-lead spike measurements and annotations to them.

In addition, it can save spike by spike measurements into one or more measurement tables. The main purpose of Section 18 ("Additional ECG Annotations") is to provide a solution for storage of any type of manually or automatically created annotations that have not been systematically stored in Sections 7, 8, 10, 11 and 15 to 17. These include: start (and end) of the original rhythm or atrial fibrillation, identification of the pacemaker spike which was not included in Part 17, and measurements not provided in section 15 and 16 (or several measurements, such as QT interval, neither section 15 nor section 16), manual annotations of more complex cases with different types of abnormal QRS complexes (LBBB aberration etc.) and P waves (AV dissociation, etc.), as well as annotations for example about the noise level in a given lead, etc.

The second standard, the *Medical Waveform Format Encoding Rules (MFER)*, in contrast to the previously described SCP-ECG standard, is used not only to encode ECG signals, but also any other medical signals. Although the MFER standard is not dedicated to use only in electrocardiography, it is in this area of interest that the most reports on its use are made [5, 12]. It is said that standard was proposed in 2004 by the Japanese standardizing organization in place of numerous signal coding methods used by manufacturers of measuring apparatus. The main task for this standard was to unify the record of raw medical data and as much as possible simplification of general-purpose medical information exchange systems. Figure 3 shows the information model used in the MFER standard. As you can see, it assumes maximum simplicity of implementation, all in order to achieve transparency of the record and wide applicability (from simple applications / devices to advanced and nested hospital information systems). As you can see, Figure 3 shows the general view on the composition of the full message packet sent in the MFER standard. Analyzing more closely Fig. 3, it can be noticed that the MFER message consists of four main groups, i.e.: descriptive information about the observations, appropriately recorded ECG waveform, information about the patient, as well as an additional description. Data on the ECG signal read are further extended to three parts, i.e.: channel information, frame information and raw data read from the electrocardiograph electrodes. In the further part of the article, a more detailed description of the frame layout regarding to the data read from the electrocardiogram. It is worth noting that the MFER standard is intended to encode only signals, while the coding of other medical data (e.g.

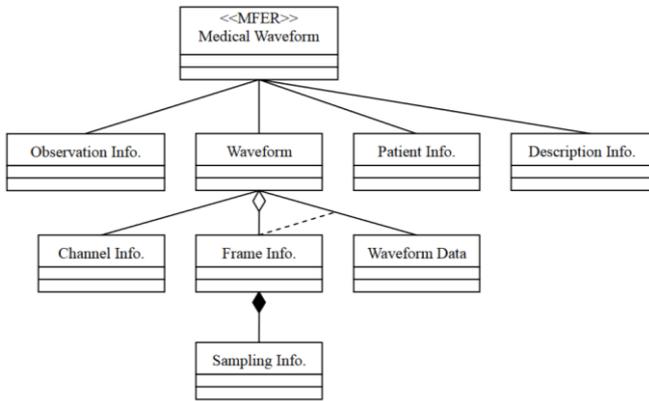


Fig. 3. The signal information model used in the MFER standard [13]

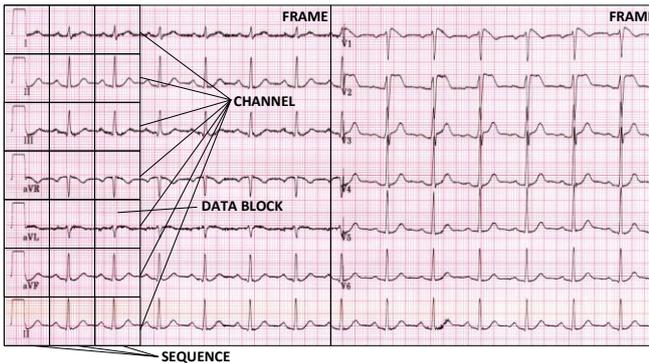


Fig. 4. Frame attributes in MFER [13]

images) that often accompany these signals MFER leaves other standards.

The MFER standard is used for coding multidimensional signals, which are divided into time frames (Figure 4). In terms of the frame, we can distinguish main attributes, which are: frame description and description of sampling. The sampling description consists of data about the frequency used and the sample resolution. The description of the frame contains information about the synchronization of signals, as well as descriptions of the three main components of the frame, i.e.: data blocks, recording channels and sequences (Figure 4).

The header and the measured signal must be coded according to the MFER rules. They should consist of type specifications (attributes such as number, primary tag and class), length specification - defining the length of the data section and value string, which is the basic information content stored in the file. The header and signal encoded in the form of MFER are shown in Figure 5.

Fig. 5 presents the general signal coding method used in the MFER standard. At this point, the question should arise how it affects the channels, sequences and data blocks, i.e. how to place data in the frame to meet the coding rules in the MFER.

To illustrate the above consideration, it is worth to transfer attention to Figure 6, which presents the detailed arrangement of the data values in the frame. As you can see in Figure 6, three channels are available (i.e. three ECG electrode leads). The recorded signal has been divided into

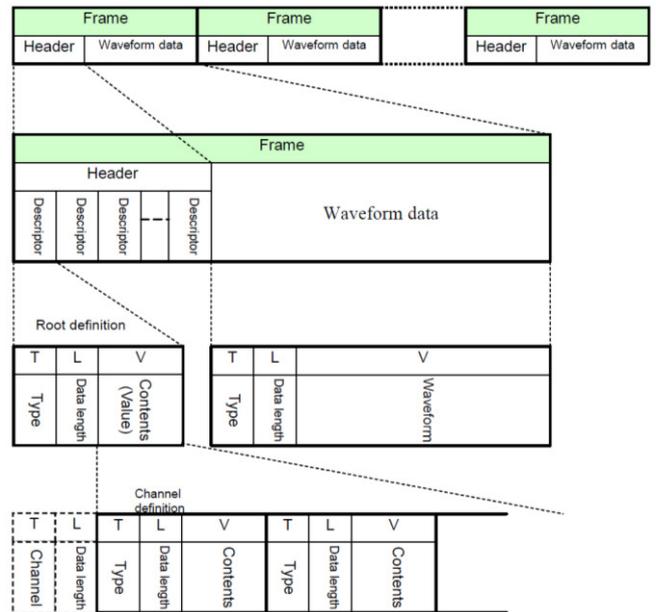


Fig. 5. Header and signal presented in the MFER encoding format [14]

four sequences, so in one frame we have to save twelve blocks of data. In addition, each block consists of five sampled signals / "beats". The result of the data arrangement in the frame will be a string consisting of a header with saved (as in Fig. 6) information about the frame and raw data from the electrodes.

For example, let suppose that sampling is done with a period of 4 ms, so the sampling frequency is 250 Hz, which gives us 20 ms for one block of recorded data (4 ms * 5 samples = 20 ms). Since there are four sequences in one frame, the waveform data for each channel is 80 ms / frame (20 ms * 4).

All complexities related to the MFER standard are widely described in the documentation available on the manufacturer's website, although not all documents are in English (a large part of the documentation is available only in Japanese).

By assumption, MFER is to meet the principle of maximum flexibility as well as the simplicity of implementation. The tags contained in the MFER have

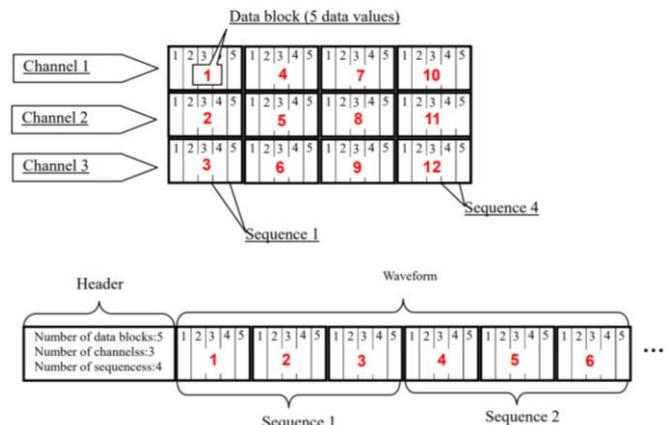


Fig. 6. Data layout in the MFER standard frame [14]

predefined default values that can be changed / adjusted when using a custom one. The MFER standard can be used to describe the signals obtained from a 12-lead electrocardiograph, a 24-hour continuous Holter recording, a vectocardiogram, stress tests, electroencephalogram, etc. Another standard that can be used to encode ECG signals is *Health Level Seven (HL7)*, and more specifically ANSI / HL7 V3 Annotated ECG (HL7 aECG). In general, HL7 can be described as the standard for electronic information exchange provided for medical environments. The standard was developed by HL7, which was established in 1987. In addition, part of the standard name, which is "level seven" can be referred to the ISO / OSI model, and more specifically the seventh layer of the model – the application layer (it can be said that HL7 uses the "transparency" of the following layers).

The Annotated ECG HL7 standard was developed in response to the initiative of the American *Food and Drug Administration (FDA)* regarding digital recording of ECG signals. A specific background to deriving this standard was the need to review and evaluate the collected ECGs in clinical trials of drugs. Until now, ECG entries with annotations were delivered to the FDA mainly in paper form, so the next logical step was to take the initiative of digitally sending ECGs with annotations in a standardized format / standard. That's why FDA, sponsors, key laboratories and manufacturers of medical devices have worked together and with HL7 to create a standard that meets all their needs.

The HL7 V3 standard is based on an object-oriented approach and the message structure is designed using the XML markup language. Each saved ECG file must be marked with a unique name (Unique ID), and should contain annotations about the stored ECG signal, which are later verified by the FDA. Due to the fact that annotations are one of the key elements of ECG recordings, this standard is known under the name Annotated ECG in XML.

Message saved in the HL7 aECG standard must be properly encoded, i.e. similar to SCP-ECG where sections are provided, here are the required and optional tags. Figure 7 shows an example of XML schema that contains a minimal description of an aECG message. A detailed description of how to implement individual tags can be found in the extensive documentation available on the manufacturer's website. Analyzing Fig. 7, which describes the absolute minimum to be met, it is clear that there is no information about the value of the recorded ECG signal. Therefore, Figure 8 shows a fragment of the XML schema, which should be added to the diagram shown in Figure 7. This fragment is responsible for storing data taken from the ECG electrode lead (in this case, lead number 1). In the first phase of the project, the recorded ECG signal was stored directly in the structure of the XML schema, and more specifically in the section described by the tag `<digits>` `</digits>`. However, there were cases where the message saved in this way took up 9 GB disk space. To overcome the problem of excessive message size, the aECG standard was slightly modified and the part strictly responsible for storing data in the XML schema was separated into a separate file. In this

```
<AnnotatedECG>
  <id root="61d1a24f-b47e-41aa-ae95-f8ac302f4eeb"/>
  <code code="93000" codeSystem="2.16.840.1.113883.6.12" codeSystemName="CPT-4"/>
  <effectiveTime>
    <center value="20021122091000"/>
  </effectiveTime>
  <componentOf>
    <timepointEvent>
      <componentOf>
        <subjectAssignment>
          <subject>
            <trialSubject>
              <id root="2.16.840.1.113883.3.456" extension="SBJ-123"/>
            </trialSubject>
          </subject>
        </componentOf>
        <clinicalTrial>
          <id root="2.16.840.1.113883.3.123" extension="PUK-123-TRL-1"/>
        </clinicalTrial>
      </componentOf>
    </subjectAssignment>
  </componentOf>
</timepointEvent>
</componentOf>
</AnnotatedECG>
```

Fig. 7. The minimum set of information that an aECG message must contain [15]

way, the raw ECG signal was transferred to a binary file, whereas in the XML schema there is only an index/pointer on the data file (example - selection in Fig. 8). Figures 7 - 9 shows the record of the ECG signal and the minimal XML schema for HL7 aECG, however, it was mentioned earlier that this is the standard of "annotations", which are further analyzed by the FDA and that is where the greatest emphasis is placed on them. The implementation documentation available on the manufacturer's website contains a number of examples of how to enter annotations, what field codes must

```
<component>
  <sequence classCode="OBS">
    <code code="MDC_ECG_LEAD_I" etc. />
    <value xsi:type="SLIST_PQ">
      <origin value="0" unit="uV"/>
      <scale value="3.75" unit="uV"/>
      <digits mediaType="application/octet-stream"
        itemType="INT"
        itemSize="2"
        headSize="0"
        recordSize="18"
        itemOffsetIntoRecord="0"
        recordCount="4320000">
        <reference value="waveforms.bin"/>
      </digits>
    </value>
  </sequence>
</component>
```

Fig. 8. Pointer on a binary file with a raw ECG signal [15]

```

<annotation classCode="OBS">
  <code codeSystemName="MDC"
codeSystem="2.16.840.1.113.883.6.24"
code="MDC_ECG_WAVC_TYPE"/>
  <value xsi:type="CE" codeSystemName="MDC"
codeSystem="2.16.840.1.113883.6.24"
code="MDC_ECG_WAVC_QRSWAVE_ONSET"/>
  <support typeCode="SPRT">
    <supportingROI classCode="ROIBND">
      <code codeSystemName="ActCode"
codeSystem="2.16.840.1.113883.5.4" code="ROIPS"/>
      <component typeCode="COMP">
        <boundary classCode="OBS">
          <code codeSystemName="ActCode"
codeSystem="2.16.1.113883.5.4"
code="TIME_ABSOLUTE"/>
          <value xsi:type="SLIST_TS">
            <origin
value="200.112.190.74300.000"/>
            <scale value="1" unit="ms"/>
            <digits mediaType="application/octet-
stream"
            itemByteOrder="LE"
            itemType="UNIT"
            itemSize="4"
            headSize="0"
            recordSize="12"
            itemOffsetIntoRecord="4"
            recordCount="4320000">
              <reference value="BeatLabels.bin"/>
            </digits>
          </value>
        </boundary>
      </component>
    </supportingROI>
  </support>
</annotation>
    
```

Fig. 9. Annotation about the QRS wave set for the HL7 aECG scheme [15]

be included in the relevant tags, and ready-made XML schemas of popular annotations. An example here is Figure 9 describing the QRS wave set. This is an example of a ready annotation that can be attached to the aECG XML schema. In addition, the selection in Figure 9 indicated a field with a code saying that we are dealing with a QRS wave and the pointer on the binary file in which the selected ECG signal was saved.

The next, listed immediately after HL7, is the standard published in 1993, and in 1995 approved in Europe *Digital Imaging and Communications in Medicine (DICOM)*. According to the mentioned name, this standard is intended for management, archiving, printing and data transmission in the field of medical imaging. Adapted on a large scale in information systems that process medical images such as x-ray pictures, magnetic resonance imaging (MR), computed tomography (CT), etc. [12]. However, according to Supplement No. 30 of the DICOM standard documentation, it can send non-pictorial data, which are related to the patient (i.e. ECG signal or other signal describing life parameters based on the recording of curves). Additionally, thanks to the introduction of time stamps, correlations between digital

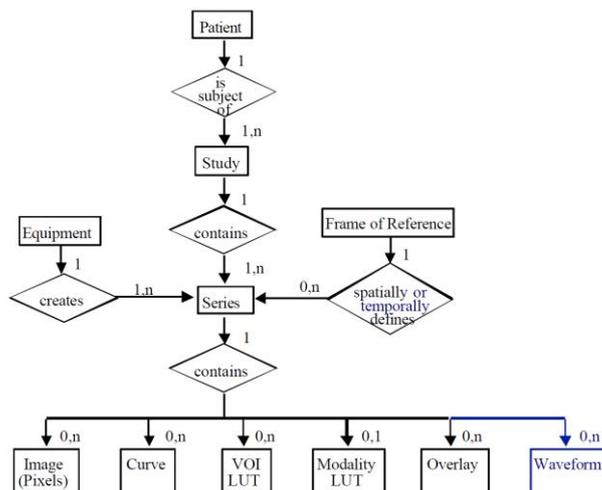


Fig. 10. Information model for data storage in the DICOM standard [16]

images can be visualized, e.g. in the angiographic examination, the heart pressure can be recorded simultaneously. Fig. 10 presents the information model for the DICOM standard, which in the final part may contain a number of different object models, including the blue model called “waveform” which contain curves (ECG or other). The specified model can be called a unit of information about the curve. Each such unit contains a number of technical parameters associated with the attributes of the recorded wave, as well as recorded samples of the raw signal. The exact structure of the curve information model is shown in Figure 11. Each such object contains information about the time in which the signal was acquired and a number of multiplex groups, each group is determined by digitizing with the same clock. The timing frequency for the clock can be defined for each multiplex group. By going down below, a group can contain one or more recording channels with a full technical definition,

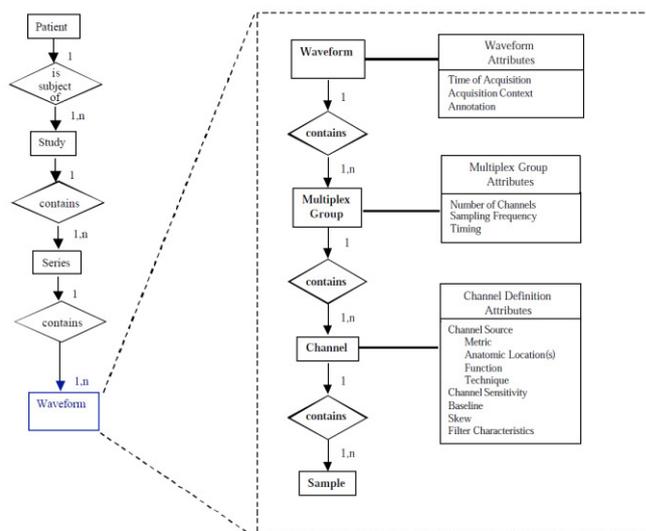


Fig. 11. Waveform information model for the DICOM standard [16]

moreover, each channel has its own separate set of digital wave samples. Additionally, when analyzing Fig. 11 it can be seen that in the "wave curve" object we can define more than one multiplex group, which gives us the opportunity to register two or more types of multi-channel signals (such as ECG and heart pressure). It is also worth mentioning that in supplement 30 of the documentation of the DICOM standard, it was specified that the definition of the "waveform" object is harmonized with the semantic structure of the HL7 standard, including the definition attributes for synchronously acquired channels. Thanks to the solution, which is based on a common object model, it is easier to share and transcode between the HL7 and DICOM standards.

VII. TELEMEDICINE CONCLUSIONS

Analyzing the specified ECG signal coding standards a specially for future use in the constructed device, it is worth focusing on the SCP-ECG standard, which is designed and used exclusively to handle signals from ECG devices. In addition, the built-in ability to compress a lossless transmitted signal can be a decisive element when using it in the field of telemedicine, or long-term measurements, where the data transmission speed or capacity of storage media device is limited. However, it is also worth considering the HL7 aECG standard, which ensures wider interoperability, and additionally the extracted file with saved raw data is an inspirational element for the application of lossless data compression algorithms.

REFERENCES

- [1] Szczeklik A., Gajewski P., Interna szczeklika 2015, Medycyna Praktyczna, Kraków, 2015
- [2] Houghton A. R., Gray D., EKG jano i zrozumiale, α -medica press, 2014
- [3] Kligfield P. and others, Recommendations for the Standardization and Interpretation of the Electrocardiogram, Journal of the American College of Cardiology Volume 49, Issue 10, 13 March 2007, Pages 1109-1127
- [4] Norma PN-EN 1064
- [5] Tadeusiewicz R. „Informatyka medyczna”, Instytut Informatyki UMCS, Lublin, 2011
- [6] Mandellos G. J., Koukias M. N., Styliadis I. S., Lymberopoulos D. K. e-SCP-ECG+ protocol: An expansion on SCP-ECG protocol for health telemonitoring—Pilot implementation. *Int. J. Telemed. Appl.* 2010;2010:1–17
- [7] Rubel P., Pani D., Schloegl A., Fayn J., Badilini F., Macfarlane P. W., Varri A. SCP-ECG V3.0: An Enhanced Standard Communication Protocol for Computer-assisted Electrocardiography, *Computing in Cardiology* 2016; 43:309-312.
- [8] <https://adst.mp.pl/img/articles/kardiologia.mp.pl/ekg/kurs/ekg-2-1.jpg>
- [9] http://www.imreference.com/_/rsrc/1467898020602/cardiology/ekg/Screen%20Shot%202016-01-18%20at%202.04.32%20PM.png
- [10] <http://webimatics.univ-lyon1.fr/scp-ecg/index.php?controler=Home&action=services>
- [11] Badilini F., The ISHNE Holter Standard Output File Format, Francia
- [12] Rune Fensli, Evaluation of international standards for ECG-recording and storage for use in tele-medical services, 2006
- [13] Medical Waveform Format Encoding Rules – Part 1 http://www.mfer.org/doc/MFER_Part1_Ver105.pdf
- [14] Medical Waveform Format Encoding Rules – Part 3-1 <http://www.mfer.org/doc/JPACS-2013-S001.pdf>
- [15] Brown B. D., Baldini F., HL7 aECG Implementation Guide Final March 21, 2005
- [16] Digital Imaging and Communications in Medicine (DICOM) Supplement 30: Waveform Interchange

11th Workshop on Computer Aspects of Numerical Algorithms

NUMERICAL algorithms are widely used by scientists engaged in various areas. There is a special need of highly efficient and easy-to-use scalable tools for solving large scale problems. The workshop is devoted to numerical algorithms with the particular attention to the latest scientific trends in this area and to problems related to implementation of libraries of efficient numerical algorithms. The goal of the workshop is meeting of researchers from various institutes and exchanging of their experience, and integrations of scientific centers.

TOPICS

- Parallel numerical algorithms
- Novel data formats for dense and sparse matrices
- Libraries for numerical computations
- Numerical algorithms testing and benchmarking
- Analysis of rounding errors of numerical algorithms
- Languages, tools and environments for programming numerical algorithms
- Numerical algorithms on coprocessors (GPU, Intel Xeon Phi, etc.)
- Paradigms of programming numerical algorithms
- Contemporary computer architectures
- Heterogeneous numerical algorithms
- Applications of numerical algorithms in science and technology

EVENT CHAIRS

- **Bylina, Beata**, Maria Curie-Skłodowska University, Poland
- **Bylina, Jaroslaw**, Maria Curie-Skłodowska University, Poland
- **Stpicyński, Przemysław**, Maria Curie-Skłodowska University, Poland

PROGRAM COMMITTEE

- **Amodio, Pierluigi**, Università di Bari, Italy

- **Anastassi, Zacharias**, De Montfort University, United Kingdom
- **Banaś, Krzysztof**, AGH University of Science and Technology, Poland
- **Brunano, Luigi**, Università di Firenze, Italy
- **Fialko, Sergiy**, Tadeusz Kościuszko Cracow University of Technology, Poland
- **Fourneau, Jean-Michel**
- **Gansterer, Wilfried**, University of Vienna, Austria
- **Georgiev, Krassimir**, IICT - BAS, Bulgaria
- **Gravvanis, George**, Democritus University of Thrace, Greece
- **Kozielski, Stanislaw**
- **Księżopolski, Bogdan**
- **Kucaba-Pietal, Anna**, Politechnika Rzeszowska, Poland
- **Lirkov, Ivan**, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria
- **Luszczek, Piotr**, University of Tennessee, United States
- **Marowka, Ami**, Bar-Ilan University, Israel
- **Petcu, Dana**, West University of Timisoara, Romania
- **Ristov, Sashko**, University of Innsbruck, Austria
- **Satco, Bianca-Renata**, Stefan cel Mare University of Suceava, Romania
- **Sergeichuk, Vladimir**, Institute of Mathematics of NAS of Ukraine, Ukraine
- **Shishkina, Olga**, Max Planck Institute for Dynamics and Self-Organization, Germany
- **Srinivasan, Natesan**, Indian Institute of Technology, India
- **Tudruj, Marek**, Inst. of Comp. Science Polish Academy of Sciences/Polish-Japanese Institute of Information Technology, Poland
- **Tůma, Miroslav**, Academy of Sciences of the Czech Republic, Czech Republic
- **Vazhenin, Alexander**, University of Aizu, Japan

Benchmarking overlapping communication and computations with multiple streams for modern GPUs

Pawel Czarnul

Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Narutowicza 11/12, 80-233 Poland
Email: pczarnul@eti.pg.edu.pl

Abstract—The paper presents benchmarking a multi-stream application processing a set of input data arrays. Tests have been performed and execution times measured for various numbers of streams and various compute intensities measured as the ratio of kernel compute time and data transfer time. As such, the application and benchmarking is representative of frequently used operations such as vector weighted sum, matrix multiplication etc. The paper shows benefits of using multiple data streams for various compute intensities compared to one stream, benchmarked for 4 GPUs: professional NVIDIA Tesla V100, Tesla K20m, desktop GTX 1060 and mobile GeForce 940MX. Additionally, relative performances are shown for various numbers of kernel computations for these GPUs.

I. INTRODUCTION

GENERAL purpose programming on GPUs (GPGPU) has become an effective approach to parallelization of many real world problems such as simulation of phenomena in 2D, 3D spaces, numerical computations, image and video processing etc. Nowadays, NVIDIA CUDA, OpenCL and OpenACC are the three dominant APIs for GPUs. Typically, several techniques are used for optimization of such programs [1]. These include, described in CUDA terms: optimization of memory referencing – global memory coalescing, proper shared memory accesses with consideration of memory banks, minimization of thread divergence, overlapping host to device communication, kernel execution and device to host communication, data prefetching from global memory to shared memory using registers, loop unrolling [2], [3]. Such changes can improve execution times significantly, at least several times and are of high importance consequently. This paper analyzes the impact of communication between the host and the device, kernel execution and device to host communication using various numbers of streams [4], for 4 various GPU models, compared to a single stream implementation for an application that can be regarded as a template useful for processing of sets of input data packets such as operations on vectors or matrices that are building blocks for many applications.

II. RELATED WORK

CUDA application and system models, numerous examples and typical aforementioned optimizations are discussed in the

literature [2], [3], also from the point of view of power/performance efficiency of different optimizations [5]. The particular problem addressed in this work can be applied to any GPU application that processes a sequence of independent input data sets for which communication and computations can be overlapped, for example a sequence of matrix multiplications, block-based matrix multiplication, computing similarities among a large number of multidimensional vectors [6] etc. Furthermore, results from this study can also be incorporated into frameworks that can automatically parallelize computations performed in batches. This is the case, for instance, for KernelHive [7] that can schedule computations and manage input and output data and run kernels on compute devices such as GPUs and CPUs. In this case the proposed technique allows overlapping communication and computations. Furthermore, the results for various compute intensities can be embedded into GPU processing models, also incorporating overlapping communication and computations, in modeling and simulation systems such as MERPSYS [8].

The matter considered in this work was found important before in the context of analysis of the optimal number of streams for best execution time in terms of the number of iterations of a loop within a kernel [9]. Analysis and a model were proposed for older cards with CUDA compute capabilities 1.x and 2.x. Practical tests, benchmarking were performed for older GPUs such as GeForce GTX 280 and GTX 480. Finally, an analytical formula was proposed to find the best number of streams. Compared to those results, we focus on current generation cards by benchmarking outcomes experimentally, plot results as relative gains compared to a single stream for various compute intensities showing minima and compare performances of 4 modern GPUs of various types – professional, desktop and mobile and not just desktop series cards.

Using multiple streams can speed up computations on a GPU or GPUs for several applications. For instance, in [10] it was shown how a multithreaded application using multiple streams increases the frame rate performance up to 78 frames per second compared to 36 for a single stream implementation for the application of real-time ultrasound elastography

suitable for diagnosis and treatment of cancer.

The authors of [11] investigated and modeled processing on a GPU using streams in the context of implementing a push based DBMS called G-SDMS. They focused especially on runtime resource scheduling using streams. Tests and verification of scheduling algorithms were performed on NVIDIA GTX680 and Tesla K40.

The authors of [12] showed how using multiple CUDA streams with multiple OpenMP threads allows to improve current state-of-the-art communication performance in an environment with multiple GPUs for a representative 3D stencil example. Tests were performed on Tesla K20, GTX590 and Tesla C2050.

In [13], the authors have presented a HCLCOOC library which enables to write data-parallel kernels for accelerators including GPUs and overlapping host-accelerator data transfer with kernel execution.

III. MODEL AND APPROACH

The application model considered in this work assumes a sequence of independent operations performed on various input data chunks. Specifically, each operation takes two data chunks as input and produces one data chunk as its output. This is depicted in Figure 1. Such a model, with such data sizes, corresponds to frequently performed operations such as vector weighted addition, matrix multiplication, blending images etc. that only differ in the ratio of computations to (host-to-device and device-to-host) communication times. In this paper, this ratio is varied in order to assess benefits of using more than 1 stream (2 and 4 tested) for various GPUs, from professional through desktop up to mobile chips. More streams allow overlapping communication and computations and consequently minimization of application execution time. Results allow to assess benefits of using multi-stream code versions for a given type of card and compute intensity of the analyzed application.

IV. EXPERIMENTS

A. Testbed environments

Within this work, we evaluate the proposed approach for 4 different, modern GPUs that differ in the target market segments:

- server data center oriented Tesla V100 and K20m,
- desktop GeForce GTX 1060,
- mobile GeForce 940MX

as well as CUDA compute capabilities:

- 7.0 – Tesla V100,
- 6.1 – GeForce GTX 1060,
- 5.0 – GeForce 940MX,
- 3.5 – Tesla K20m.

This means that the tests can be considered as representative both in terms of GPU types and compute capabilities. Specific parameters of the GPUs are listed in Table I.

B. Tests and results

For each of the GPUs, tests were performed for various compute intensities and various numbers of streams: 1, 2 and 4. Additionally, two ways of issue orders were tested for 2 and 4 streams:

- order A – $n \times \{ \text{host to device copy, kernel launch, device to host copy} \}$,
- order B – $n \times \text{host to device copy, } n \times \text{kernel launch, } n \times \text{device to host copy}$.

Compute intensity is defined as the ratio of kernel compute time for each data packet divided by each data packet size, measured for 1 stream. We should note that such an application and various compute intensities are representative of many real applications because:

- 1) There are many applications that take input data of size $2n$ and produce output of size n . Typical examples include matrix multiplication and vector addition, used in numerous codes. Notable applications using such operations nowadays include, for example, machine learning [14], weather pattern analysis [15], shortest path problem etc. [16].
- 2) Various compute intensities correspond to various operations such as: higher for matrix multiplication, lower for addition of vectors etc.

Figures 2 through 9 present comparison of execution times obtained for various compute intensities for all the GPUs tested, expressed both in terms of actual execution times as well as relative times compared to 1 stream versions.

From the latter we can easily see distinct features that impact potential uses of these results for actual applications:

- Except the 940MX, there is about 20-30% gain in using 2, 4 streams compared to 1 stream for small compute intensities (<1).
- There is an increase of gain from using 2, 4 streams up to around 50%, except the 940MX for which the gain reaches around 35%.
- Gain from using 2, 4 streams diminishes as the compute intensity grows.
- There is a gain from using 4 streams compared to 2 streams for compute intensities up to around 2-4. The average gains are as follows for particular GPUs:
 - Tesla K20m – 4.9%,
 - GTX 1060 – 3.5%,
 - GeForce 940MX – 3.3%,
 - Tesla V100 – 4.5%.

Minima observed are due to the fact that too low and too high compute intensities do not allow considerable overlapping of communication and computations.

In terms of comparisons of issue orders, Tables II and III present differences between the aforementioned B and A orders in percentages, out of minimums out of 5 runs for each configuration. It can be seen that differences for all the GPU tested do not exceed 1.4% which means that are negligible. Configurations with the same number of computations in the

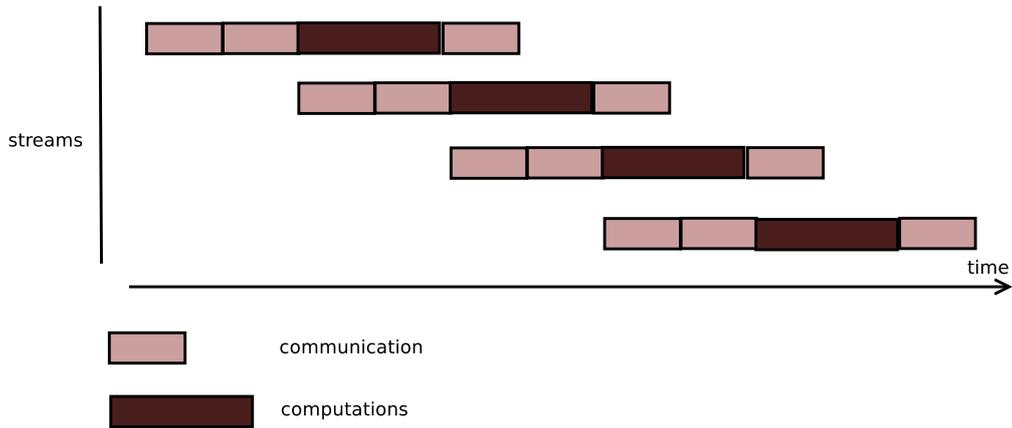


Figure 1. Streams

Table I
GPUS – SPECIFICATIONS

GPU	Tesla V100	Tesla K20m	GTX 1060	GeForce 940MX (GDDR5)
Architecture name	Volta	Kepler	Pascal	Maxwell
CUDA compute capability	7.0	3.5	6.1	5.0
CUDA cores	5120	2496	1280	512
clock [MHz]	1455	706	1708	861
memory size [GB]	16	5	6	4
memory bus width [bits]	4096	320	192	64
memory bandwidth [GB/s]	900	208	192	40.1
power [W]	300	225	120	23

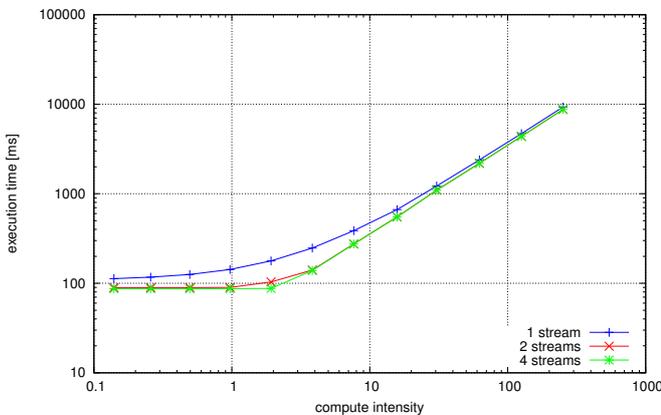


Figure 2. Execution times on Tesla V100

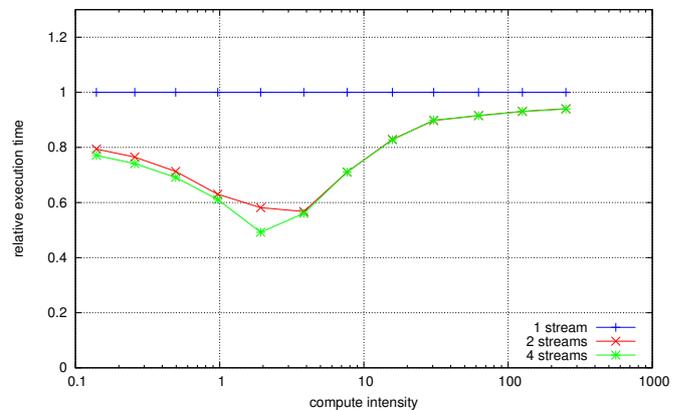


Figure 3. Relative execution times on Tesla V100

kernel (called relative computation count, not to be confused with the assumed definition of compute intensity) were compared against each other for all the GPUs tested.

Furthermore, we have compared relative performance per data size of the GPUs, proportional to the inverse of application execution time, for the same relative computation counts. Performances were scaled against the smallest configuration on the slowest GPU out of the tested ones – GeForce 940MX. Results are shown in Figure 10. It can be seen that perfor-

mances reach their maximum for certain computation counts and stay such for larger computation counts. Relative order of the GPUs is, from the fastest: V100, GTX 1060, Tesla K20m and 940MX. It should be noted that relative performances may depend on the kernel code, which in this case computes a weighted sum of two input vectors, with the number of computations that can be defined for testing various compute intensities.

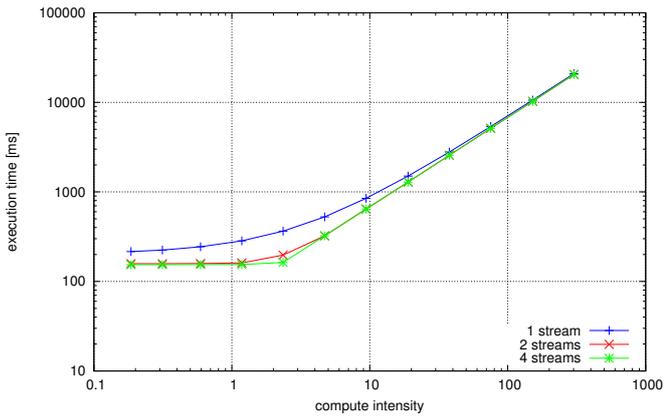


Figure 4. Execution times on Tesla K20m

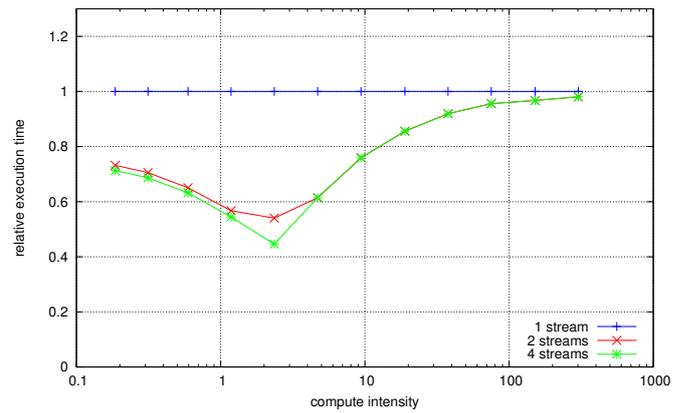


Figure 5. Relative execution times on Tesla K20m

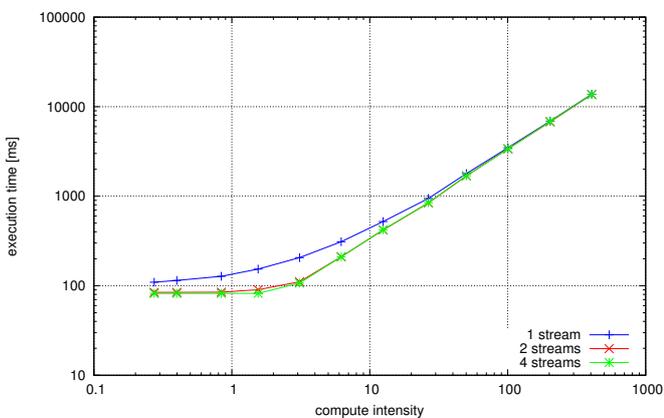


Figure 6. Execution times on GTX 1060

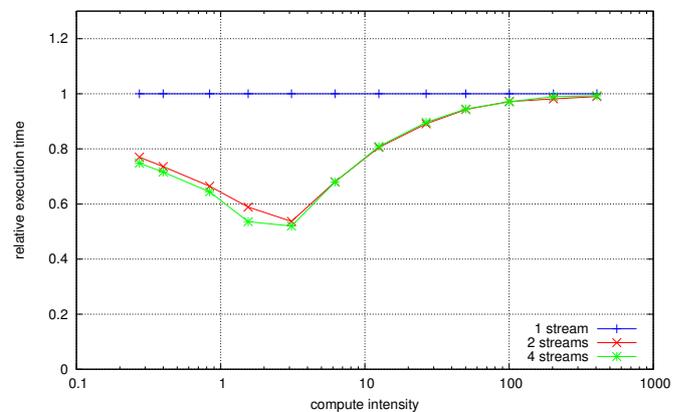


Figure 7. Relative execution times on GTX 1060

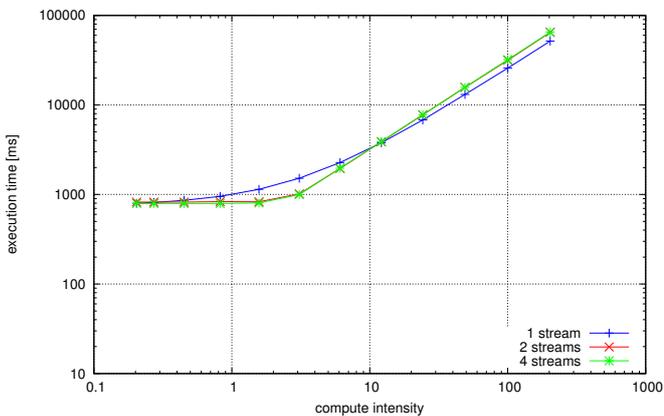


Figure 8. Execution times on GF 940MX

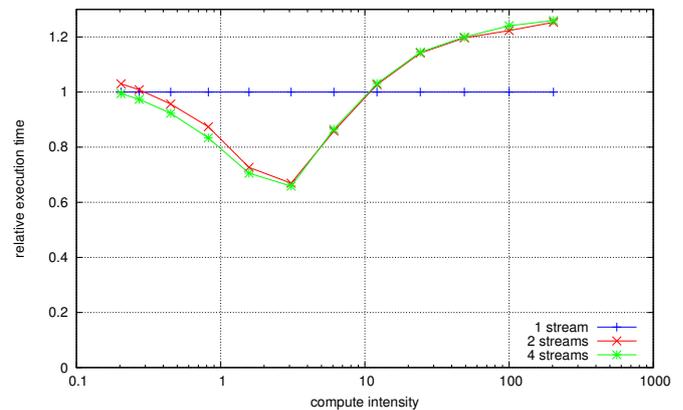


Figure 9. Relative execution times on GF 940MX

V. SUMMARY AND FUTURE WORK

In the paper, we have compared execution times of an application processing multiple data chunks on a GPU, in several versions that differ in the number of CUDA streams used. As expected using 2 and 4 streams brings benefits compared to 1 stream. A non-trivial maximum of gains out of using

multiple streams has been shown in terms of various compute intensities for 4 GPUs tested including the latest professional NVIDIA V100, desktop GTX 1060, professional Tesla K20m and mobile GeForce 940MX. It has been shown that gain from using 4 streams compared to 2 streams is visible up to compute intensities of around 2-4. Additionally, performances

Table II
B VS A ISSUE ORDER DIFFERENCE [%] – 2 STREAMS

relative computation count	Tesla V100	Tesla K20m	GTX 1060	GeForce 940MX
1.00	0.00	0.06	0.00	-0.04
2.00	-0.45	0.06	0.00	0.00
4.00	-0.11	0.13	-0.24	0.01
8.00	-0.34	0.37	-0.11	-0.35
16.00	-0.22	0.36	0.09	0.00
32.00	0.00	0.25	0.00	-0.01
64.00	0.11	0.12	-0.05	0.00
128.00	-0.10	0.05	-0.23	0.03
256.00	-0.07	0.04	-0.02	0.17
512.00	0.40	0.00	-0.04	0.12
1024.00	-0.22	0.01	-0.01	0.76
2048.00	0.14	0.01	0.00	0.56

Table III
B VS A ISSUE ORDER DIFFERENCE [%] – 4 STREAMS

relative computation count	Tesla V100	Tesla K20m	GTX 1060	GeForce 940MX
1.00	0.00	-0.07	0.12	-0.06
2.00	0.00	-0.06	-0.12	-0.04
4.00	-0.12	-0.06	-0.12	-0.05
8.00	-0.23	-0.06	0.36	-0.13
16.00	0.12	0.49	-0.56	-1.39
32.00	0.23	0.22	0.33	-1.14
64.00	0.11	0.12	-0.14	-0.60
128.00	-0.11	0.05	-0.22	-0.15
256.00	-0.58	0.03	0.04	0.42
512.00	0.47	0.02	0.52	0.28
1024.00	-0.11	0.00	-0.52	0.29
2048.00	0.12	0.00	-0.04	0.73

of the cards were compared showing both relative values as well as how performance per data size grows for each card for growing numbers of computations in the GPU kernel. Tesla V100 outperforms the other cards significantly reaching its top performance for higher relative computation count. It is interesting to see that the current desktop series GTX 1060 outperforms visibly a few years old professional series Tesla K20m. The mobile GPU cannot match the other cards in performance but its performance grows for larger relative computation counts compared to Tesla K20m and GTX 1060.

In the future, the author plans to extend the set of experiments in terms of the following: testing various kernel codes on various devices and how these impact the ratio of computations to communication, testing particular kernels operating on data types of various precision as this impacts the ratio of computations to communication and finally focusing on performance/real power consumption for the tested GPUs.

ACKNOWLEDGMENTS

Tests partially performed on Tesla V100 within DGX station purchased and installed at the Faculty of Electronics,

Telecommunications and Informatics, Gdansk University of Technology, Poland.

Work was supported partially by the Polish Ministry of Science and Higher Education.

REFERENCES

- [1] C. Woolley, "Gpu optimization fundamentals," February 2013, nVIDIA Developer Technology Group, https://www.olcf.ornl.gov/wp-content/uploads/2013/02/GPU_Opt_Fund-CW1.pdf.
- [2] J. Sanders and E. Kandrot, *CUDA by Example: An Introduction to General-Purpose GPU Programming*, 1st ed. Addison-Wesley Professional, 2010. ISBN 0131387685, 9780131387683
- [3] P. Czarnul, *Parallel Programming for Modern High Performance Computing Systems*. CRC Press, 2018, ISBN 9781138305953.
- [4] J. Luitjens, "Cuda streams. best practices and common pitfalls," 2014, nVIDIA, <http://on-demand.gputechconf.com/gtc/2014/presentations/S4158-cuda-streams-best-practices-common-pitfalls.pdf>.
- [5] Y. Ukidave, A. K. Ziabari, P. Mistry, G. Schirner, and D. Kaeli, "Analyzing power efficiency of optimization techniques and algorithm design methods for applications on heterogeneous platforms," *The International Journal of High Performance Computing Applications*, vol. 28, no. 3, pp. 319–334, 2014. doi: 10.1177/1094342014526907. [Online]. Available: <https://doi.org/10.1177/1094342014526907>
- [6] P. Czarnul, "Parallelization of large vector similarity computations in a hybrid cpu+gpu environment," *The Journal of Supercomputing*,

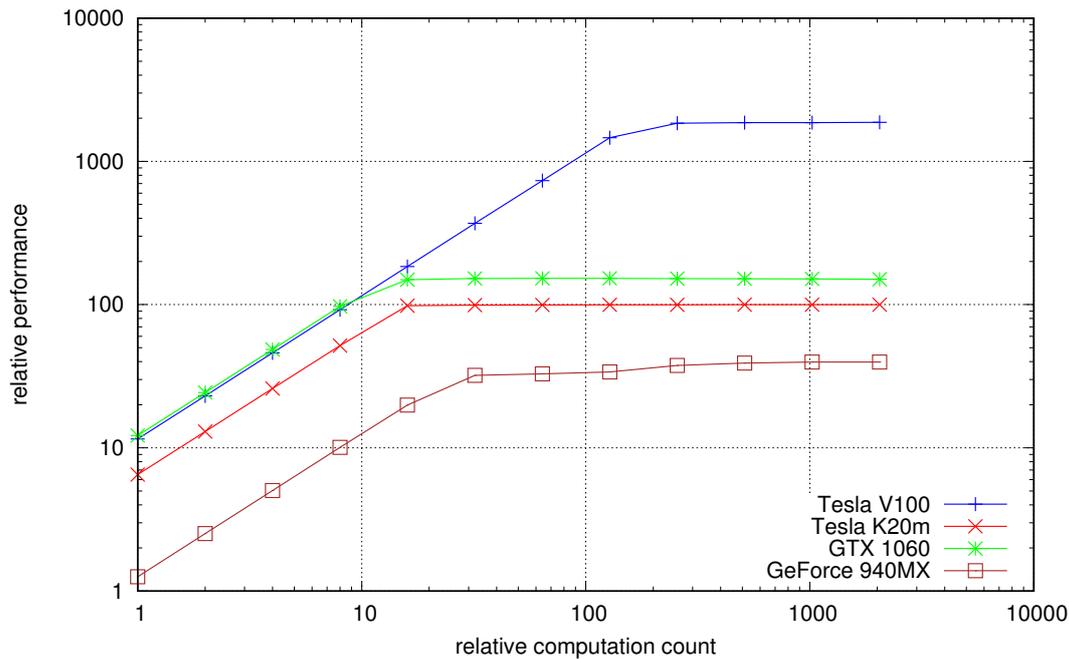


Figure 10. Relative performance vs relative computation count

- vol. 74, no. 2, pp. 768–786, Feb 2018. doi: 10.1007/s11227-017-2159-7. [Online]. Available: <https://doi.org/10.1007/s11227-017-2159-7>
- [7] P. Rościszewski, P. Czarnul, R. Lewandowski, and M. Schally-Kacprzak, “Kernelhive: a new workflow-based framework for multilevel high performance computing using clusters and workstations with cpus and gpus,” *Concurrency and Computation: Practice and Experience*, vol. 28, no. 9, pp. 2586–2607. doi: 10.1002/cpe.3719. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.3719>
- [8] P. Czarnul, J. Kuchta, M. Matuszek, J. Proficz, P. Rościszewski, M. Wójcik, and J. Szymański, “Merpsys: An environment for simulation of parallel application execution on large scale hpc systems,” *Simulation Modelling Practice and Theory*, vol. 77, pp. 124 – 140, 2017. doi: <https://doi.org/10.1016/j.simpat.2017.05.009>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1569190X17300916>
- [9] J. Gómez-Luna, J. M. González-Linares, J. I. Benavides, and N. Guil, “Performance models for asynchronous data transfers on consumer graphics processing units,” *Journal of Parallel and Distributed Computing*, vol. 72, no. 9, pp. 1117 – 1126, 2012. doi: <https://doi.org/10.1016/j.jpdc.2011.07.011> Accelerators for High-Performance Computing. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0743731511001468>
- [10] N. P. Deshmukh, H. J. Kang, S. D. Billings, R. H. Taylor, G. D. Hager, and E. M. Boctor, “Elastography using multi-stream gpu: an application to online tracked ultrasound elastography, in-vivo and the da vinci surgical system,” *PloS one*, vol. 9, no. 12, p. e115881, 2014. doi: 10.1371/journal.pone.0115881. [Online]. Available: <http://europepmc.org/articles/PMC4277422>
- [11] H. Li, D. Yu, A. Kumar, and Y. C. Tu, “Performance modeling in cuda streams – a means for high-throughput data processing,” in *2014 IEEE International Conference on Big Data (Big Data)*, Oct 2014. doi: 10.1109/BigData.2014.7004245 pp. 301–310.
- [12] M. Sourouri, T. Gillberg, S. B. Baden, and X. Cai, “Effective multi-gpu communication using multiple cuda streams and threads,” in *2014 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS)*, Dec 2014. doi: 10.1109/PADSW.2014.7097919. ISSN 1521-9097 pp. 981–986.
- [13] H. Khaleghzadeh, Z. Zhong, R. Reddy, and A. Lastovetsky, “Out-of-core implementation for accelerator kernels on heterogeneous clouds,” *The Journal of Supercomputing*, vol. 74, no. 2, pp. 551–568, Feb 2018. doi: 10.1007/s11227-017-2141-4. [Online]. Available: <https://doi.org/10.1007/s11227-017-2141-4>
- [14] K. Osawa, A. Sekiya, H. Naganuma, and R. Yokota, “Accelerating matrix multiplication in deep learning by using low-rank approximation,” in *2017 International Conference on High Performance Computing Simulation (HPCS)*, July 2017. doi: 10.1109/HPCS.2017.37 pp. 186–192.
- [15] V. Yegnanarayanan, “An application of matrix multiplication,” *Resonance*, vol. 18, no. 4, pp. 368–377, Apr 2013. doi: 10.1007/s12045-013-0052-0. [Online]. Available: <https://doi.org/10.1007/s12045-013-0052-0>
- [16] W. Liu and B. Vinter, “A framework for general sparse matrix-matrix multiplication on gpus and heterogeneous processors,” *CoRR*, vol. abs/1504.05022, 2015. [Online]. Available: <http://arxiv.org/abs/1504.05022>

Accelerating Minimum Cost Polygon Triangulation Code with the TRACO Compiler

Marek Palkowski, Włodzimierz Bielecki

West Pomeranian University of Technology in Szczecin

ul. Żołnierska 49, 71-210 Szczecin, Poland

Email: mpalkowski@wi.zut.edu.pl, wbielecki@wi.zut.edu.pl

Abstract—In this paper, we present automatic loop tiling and parallelization for the minimum cost polygon triangulation (MCPT) task. For this purpose, we use the authorial source-to-source TRACO compiler. MCPT is a recursive algorithm encountering each subproblem many times in different branches of its recursion tree. The most intensive computing part is a triple nested polyhedral program loop nest filling a cost table using the MCPT recursive. First, the code is tiled by means of the transitive closure of a dependence graph. TRACO allows for tiling of the innermost loop nest that is not possible by means of other closely related compilers. We tile only the two innermost loops and apply skewing to serialize the outermost one and parallelize the innermost ones. An experimental study carried out on multi-core computers demonstrates considerable speed-up of tiled code, which overcomes that obtained for code generated with the closely related P_{Lu}To compiler based on the affine transformations framework.

I. INTRODUCTION

The cost of moving data from main memory can be higher than the cost of computation on modern multi-core platforms. This disparity between communication and computation prompts to design algorithms for better locality and parallelism. Loop nest tiling allows for both coarsening code parallelism and improving its locality that leads to increasing parallel code performance. Widely-known tiling techniques based on the polyhedral model¹ use linear or affine transformations of program loop nests [2], [3], [4], [5], [6].

Dynamic programming (DP) is typically applied to optimization problems. In such problems, there can be many possible solutions and we wish to find a solution with the optimal (minimum or maximum) value. Computing intensive DP tasks like minimal cost polygon triangulation can be presented as loop nests within the polyhedral model, however, they involve non-uniform dependences. This limits many optimization techniques such as permutation, diamond tiling [7], or index set splitting [8] to improve cache efficiency.

State-of-the-art automatic optimizing compilers, such as P_{Lu}To [2], have provided empirical confirmation of the success of polyhedral-based optimization. P_{Lu}To and similar optimizing compilers apply the affine transformation framework (ATF), which has demonstrated considerable success in generating high-performance parallel codes in particular for

¹The polyhedral model is a mathematical formalism for analyzing, parallelizing, and transforming program loop nests whose all bounds and all conditions are affine expressions in the loop iterators and symbolic constants called parameters [1].

stencils. However, in general, this framework is not able to tile all loops in dynamic programming code [9], [10].

In this paper, we use an alternative approach based on the transitive closure of dependence graphs, which allows us to tile bands of non-permutable loops [11] and extract parallelism when affine transformations miss it [11]. This approach is implemented in the TRACO [12] compiler.

TRACO does not find and use any affine function to transform the loop nest. It is based on the Iteration Space Slicing Framework introduced by Pugh and Rosser [13] and applies the transitive closure of a dependence graph to carry out corrections of original rectangular tiles so that all dependences available in the original loop nest are preserved under the lexicographic order of target tiles. The transitive closure of a graph G is a graph where there is an edge between vertices if they are connected directly or indirectly in the graph G .

In this paper, we show that such a technique enables tiling for all MCPT loops in opposite to affine transformation algorithms implemented in P_{Lu}To. We discuss the performance of parallel tiled MCPT code generated by TRACO and executed on modern multi-core processors and compare it with that of P_{Lu}To tiled code.

II. MINIMAL COST POLYGON TRIANGULATION

A polygon is a piecewise linear closed curve in the plane. A convex polygon has interior angles that are each strictly less than 180° . A triangulation of a polygon is a set of chords of the polygon that divide the polygon into disjoint triangles (polygons with 3 sides).

In the optimal (polygon) triangulation problem, we are given a convex polygon and a weight function defined on triangles formed by sides and chords. The problem is to find a triangulation that minimizes the sum of the weights of the triangles in the triangulation.

Let $\text{cost } w(i, j, k)$ denote the length of the perimeter of $\Delta v_i v_j v_k = |v_i v_j| + |v_j v_k| + |v_k v_i|$. Then minimal cost polygon triangulation is as follows,

$$c[i][j] = \begin{cases} 0 & j < i + 2, \\ \max_{i < k < j} (c[i][k] + c(k)[j] + w(i, j, k)) & \text{otherwise.} \end{cases} \quad (1)$$

Possible values of $c[i][j]$ fall into two cases. If $j < i + 2$ then the polygon with vertices $v_i \dots v_j$ has fewer than 3 vertices,

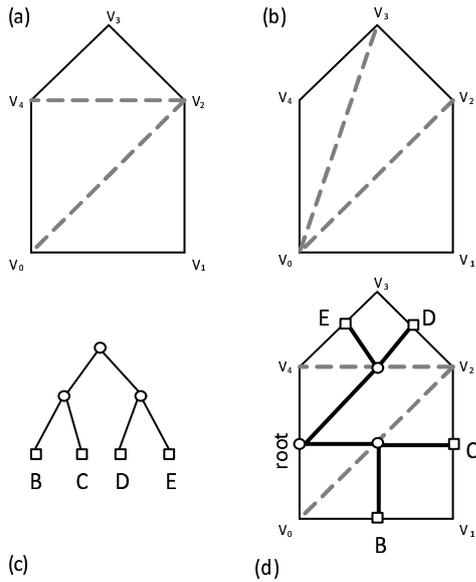


Fig. 1. The example of minimal cost polygon triangulation

and no triangulation is possible, so an appropriate minimum triangulation cost is 0, otherwise there is one or more choices of k where $i < k < j$. To list all triangles, we have to traceback using a history recorded in a separate array of the best vertex to do recursive triangulations at each step.

The memory complexity of the defined cost matrix is $\mathcal{O}(n^2)$. The time complexity of a direct implementation of this algorithm is $\mathcal{O}(n^3)$.

Figures 1a and 1b present two choices of given polygon triangulation (except mirror ones). There is a surprising correspondence between the triangulation of a polygon and the parenthesization of an expression. This correspondence is best explained using trees, see Figure 1c. A full parenthesization of an expression corresponds to a full binary tree, sometimes called the parse tree of an expression. Each leaf of a parse tree is labeled by one of the polygon sides. The root of a tree is a side between the first and last vertices. The parse tree for the parenthesized product is defined with the expression $((BC)(DE))$. The triangulation of the polygon with the parse tree overlaid is depicted in Figure 1d (assuming that triangulation presented in Figure 1a has the minimal cost).

Summing up, minimal cost polygon triangulation corresponds to dynamic programming tasks like chain matrix multiplication or an optimal binary search tree.

III. LOOP TILING AND PARALLELIZATION OF MCPT CODE

To find the minimal cost for cell $c[i][j]$, previous cells are scanned in the corresponding row and column. This is typical for dynamic programming tasks such as Nussinov's algorithm [14]. The MCPT algorithm is also within nonserial polyadic dynamic programming (NPDP). The term nonserial polyadic stands for another family of dynamic programming (DP) with nonuniform data dependences (some elements of dependence

distance vector are not constant), which is more difficult to be optimized.

Our idea to form valid target tiles is different from that based on affine transformations. First, we apply the transitive closure of the dependence graph representing all the dependences available in the loop nest, to check whether the original tiles are valid. A valid tile with identifier II does not include any dependence destination whose corresponding dependence source belongs to a tile whose identifier is greater than II . If there exist invalid tiles, we correct them with transferring invalid destinations to the tiles including the corresponding sources [11]. It is worth noting that there is no cycle in the corresponding inter-tile dependence graph and any parallelization technique can be applied.

Listing 1 presents polyhedral affine loop nest calculating the cost table of polygon triangulation defined using formula (Eq. 1). It is worth noting that the table is filled in a diagonal fashion, i.e., from diagonal elements to element $[0][n-1]$.

The loop nest can be tiled by both P_LU_To and TRACO, however, only TRACO allows us to tile all the three loops of the nest. We discovered empirically that the best tile size is $[1 \times 128 \times 16]$, i.e., the first loop has not been tiled. The second loop is parallel, it does not carry any dependence because cells are scanned in a diagonal fashion [14]. Listing 2 presents the tiled parallel OpenMP code generated by TRACO. The compiler automatically detects that the second loop enumerating tiles does not carry any dependence.

Such a code cannot be generated by means of P_LU_To because it is able to tile only the two outermost loops, the innermost loop remains untiled. For tiled code generated by P_LU_To, we empirically discover that the best tile size is $[8 \times 8 \times 1]$, P_LU_To code can be found at <https://sourceforge.net/p/traco/code/HEAD/tree/trunk/examples/trian.c>.

IV. EXPERIMENTAL STUDY

This section presents the results of the comparison of TRACO and P_LU_To tiled code performance. To carry out experiments, we have used a computer with the following features: Intel Xeon CPU E5-2699 v2, 3.6GHz, 18 cores, 36 Threads, 45 MB Cache, 16 GB RAM. We examined parallel code performance also using a coprocessor Intel Xeon Phi 7120P (16GB, 1.238 GHz, 61 cores, 30.5 MB Cache). Programs were compiled with the Intel C Compiler (icc 15.0.2) and optimized at the $-O3$ level.

Table 1 presents execution times (in seconds) for various numbers of random points of polygon vertices. Figure 2 depicts the speed-up and efficiency of the tiled programs. Analyzing the results obtained, we may conclude that the TRACO code performance overcomes that of the P_LU_To one. For one and two threads, super-linear speed-up is observed. Tiling of the innermost loop allows us to achieve the minimal execution time even without using all threads available on the computer used for experiments. Although P_LU_To code seems to be more scalable regarding the number of threads, poor locality limits its speed-up on the modern multi-core machine used for experiments.

Listing 1. Serial loop nest implementing minimum cost polygon triangulation.

```

1  for (gap = 0; gap < N; gap++){
2    for (j = gap; j < N; j++){ // i = j - gap
3      if (gap < 2) // polygon vi...vj has fewer than 3 vertices,
4         table[j-gap][j] = 0;
5      else{
6         table[j-gap][j] = INT_MAX;
7         for (k = j-gap+1; k < j; k++){
8           table[j-gap][j] = MIN(table[j-gap][j], table[j-gap][k] + ↵
           ↵ table[k][j] + cost(j-gap,j,k));
9        } } } }

```

Listing 2. Parallel tiled loop nest implementing minimum cost polygon triangulation.

```

1  for( c1 = 0; c1 < N; c1 += 1) // tiles of gap (serial)
2    #pragma omp parallel for
3    for( c3 = 0; c3 <= (N - c1 - 1) / 128; c3 += 1) { // tiles of j (parallel)
4      if (c1 >= 2) {
5        for( c4 = 1; c4 <= 2; c4 += 1) {
6          if (c4 == 2) {
7            for( c5 = 0; c5 <= floord(c1 - 2, 16); c5 += 1) // tiles of k (serial)
8              for( c9 = c1 + 128 * c3; c9 <= min(N-1, c1 + 128 * c3 + 127); c9 += 1)
9                for( c11 = -c1 + 16 * c5 + c9 + 1; c11 <= min(c9 - 1, -c1 + 16 * ↵
10                  ↵ c5 + c9 + 16); c11 += 1)
11                  table[c9-c1][c9] = MIN(table[c9-c1][c9], table[c9-c1][c11] + ↵
12                  ↵ table[c11][c9] + cost[c9-c1][c9][c11]);
13          } else
14            for( c9 = c1 + 128 * c3; c9 <= min(N - 1, c1 + 128 * c3 + 127); c9 += 1)
15              table[c9-c1][c9] = INT_MAX;
16        } } else
17          for( c9 = c1 + 128 * c3; c9 <= min(N - 1, c1 + 128 * c3 + 127); c9 += 1)
18            table[c9-c1][c9] = 0;
19    } }

```

V. CONCLUSION

In this paper, we presented the usage of the TRACO compiler, implementing optimizing loop nest algorithms based on the transitive closure of dependence graphs, to the dynamic programming of minimal cost polygon triangulation. Results of experiments demonstrate that the speed-up of parallel tiled code generated by TRACO is higher than that of code generated by means of the state-of-the-art PLuTo compiler. Tiling the innermost loop of the examined loop nest allows us to considerably accelerate NDPD programs.

In future, we plan to study arbitrarily shaped tiling based on transitive closure aimed at generating more flexible code for affine loop nests typical for a code of dynamic programming code.

REFERENCES

- [1] W. Kelly and W. Pugh, "A framework for unifying reordering transformations," Univ. of Maryland Institute for Advanced Computer Studies Report No. UMIACS-TR-92-126.1, College Park, MD, USA, Tech. Rep., 1993.
- [2] U. Bondhugula *et al.*, "A practical automatic polyhedral parallelizer and locality optimizer," *SIGPLAN Not.*, vol. 43, no. 6, pp. 101–113, Jun. 2008. doi: 10.1145/1379022.1375595 [Http://pluto-compiler.sourceforge.net/](http://pluto-compiler.sourceforge.net/).
- [3] M. Griebl, "Automatic parallelization of loop programs for distributed memory architectures," 2004.
- [4] F. Irigoien and R. Triolet, "Supernode partitioning," in *Proceedings of the 15th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, ser. POPL '88. New York, NY, USA: ACM, 1988. doi: 10.1145/73560.73588. ISBN 0-89791-252-7 pp. 319–329.
- [5] A. Lim, G. I. Cheong, and M. S. Lam, "An affine partitioning algorithm to maximize parallelism and minimize communication," in *In Proceedings of the 13th ACM SIGARCH International Conference on Supercomputing*. ACM Press, 1999. doi: 10.1145/305138.305197 pp. 228–237.
- [6] J. Xue, "On tiling as a loop transformation," 1997.
- [7] U. Bondhugula, V. Bandishti, and I. Pananilath, "Diamond tiling: Tiling techniques to maximize parallelism for stencil computations," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 5, pp. 1285–1298, May 2017. doi: 10.1109/tpds.2016.2615094
- [8] U. Bondhugula, A. Acharya, and A. Cohen, "The pluto+ algorithm: A practical approach for parallelization and locality optimization of affine loop nests," *ACM Trans. Program. Lang. Syst.*, vol. 38, no. 3, pp. 12:1–12:32, Apr. 2016. doi: 10.1145/2896389

TABLE I
EXECUTION TIME (IN SECONDS) OF THE ORIGINAL, TRACO AND PLUTo TILED CODES IMPLEMENTING MCPT.

N	1 Thread			2 Threads		4 Threads		8 Threads		16 Threads		32 Threads	
	Orig.	TRACO	PLuTo	TRACO	PLuTo	TRACO	PLuTo	TRACO	PLuTo	TRACO	PLuTo	TRACO	PLuTo
1000	2.22	1.17	1.75	0.85	1.30	0.68	1.15	0.67	01.08	0.61	0.98	0.58	0.78
1500	7.14	3.21	6.42	2.37	4.71	2.25	3.48	1.97	2.75	2.16	2.66	2.31	2.75
2000	16.78	7.19	15.42	5.58	9.67	4.61	7.89	4.75	6.98	4.76	5.91	4.78	6.33
2500	34.93	14.14	33.40	9.97	20.98	10.73	17.96	10.42	15.14	10.25	13.24	11.03	11.55
3000	62.06	24.61	63.54	16.94	42.09	16.32	32.31	17.92	28.88	17.73	21.16	17.23	20.99
5000	524.04	197.48	465.38	94.37	266.62	83.28	186.24	87.54	137.11	75.67	118.73	87.02	96.29

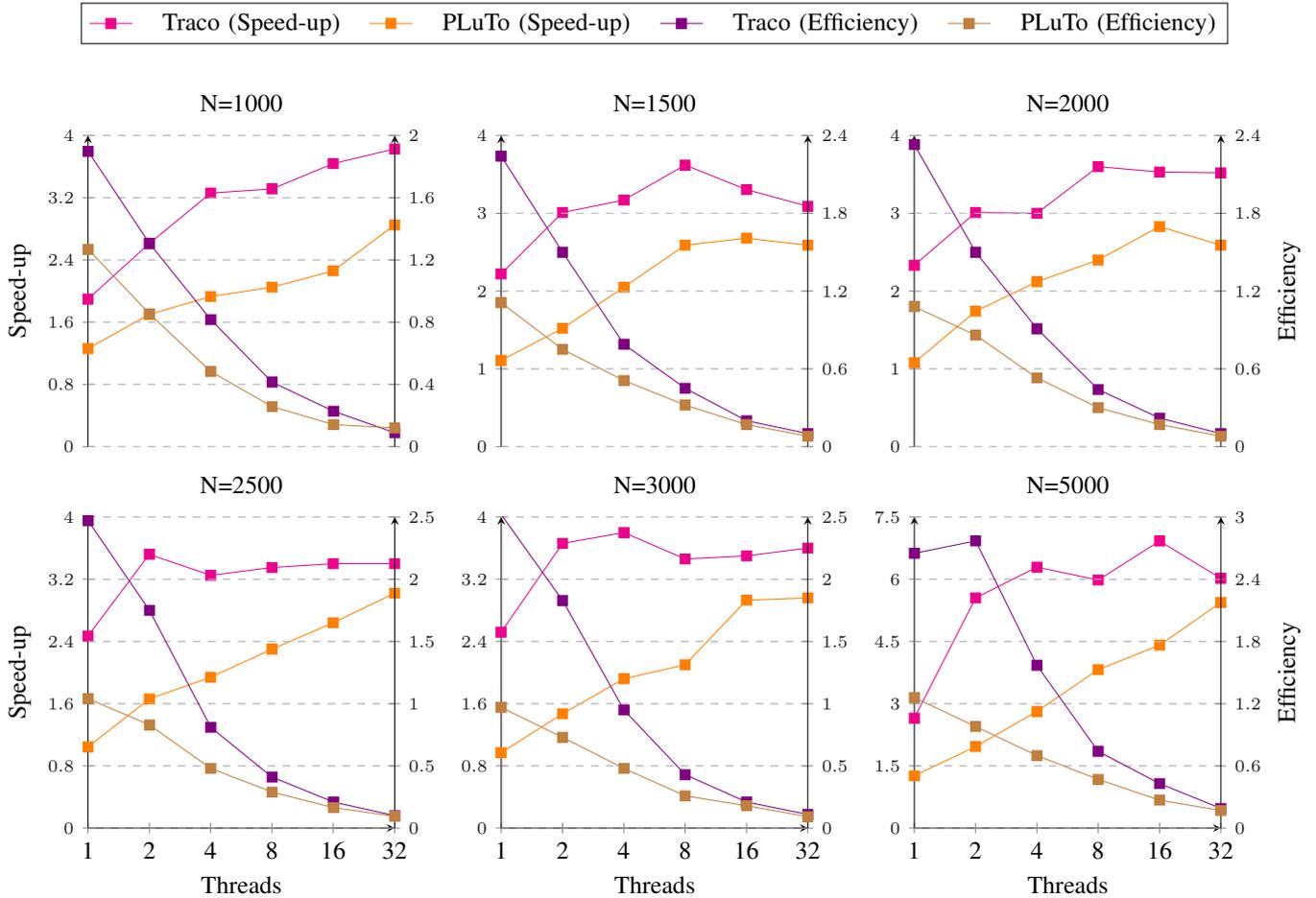


Fig. 2. Speed-up and efficiency of the TRACO and PLuTo codes.

- [9] D. G. Wonnacott and M. M. Strout, "On the scalability of loop tiling techniques," in *Proceedings of the 3rd International Workshop on Polyhedral Compilation Techniques (IMPACT)*, January 2013.
- [10] R. T. Mullapudi and U. Bondhugula, "Tiling for dynamic scheduling," in *Proceedings of the 4th International Workshop on Polyhedral Compilation Techniques*, Vienna, Austria, Jan. 2014.
- [11] W. Bielecki and M. Palkowski, "Tiling of arbitrarily nested loops by means of the transitive closure of dependence graphs," *International Journal of Applied Mathematics and Computer Science (AMCS)*, vol. 26, no. 4, pp. 919–939, December 2016. doi: 10.1515/amcs-2016-0065
- [12] —, "A parallelizing and optimizing compiler - traco," 2013. [Online]. Available: <http://traco.sourceforge.net>
- [13] W. Pugh and E. Rosser, "Iteration space slicing and its application to communication optimization," in *International Conference on Supercomputing*, 1997. doi: 10.1145/263580.263637 pp. 221–228.
- [14] M. Palkowski and W. Bielecki, "Parallel tiled Nussinov RNA folding loop nest generated using both dependence graph transitive closure and loop skewing," *BMC Bioinformatics*, vol. 18, no. 1, p. 290, 2017. doi: 10.1186/s12859-017-1707-8

Parallelizing the code of the Fokker-Planck equation solution by stochastic approach in Julia programming language

Anna Wawrzynczak*,†

*Institute of Computer Sciences, Siedce Univeristy
ul.3 Maja 54, Siedlce, Poland
e-mail: awawrzynczak@uph.edu.pl

†National Centre for Nuclear Research
ul A.Soltana 7, Świerk-Otwock, Poland

Abstract—Presenting a reliable physical simulation requires very often use of the supercomputers and models run for many days or weeks. The numerical computing is divided into two groups. One uses highly efficient low-level languages like Fortran, C, and C++. The second applies high-level languages like Python or Matlab, being usually quite slow in HPC applications. This paper presents the application of the relatively new programming language Julia, advertised as the as "a high-level, high-performance dynamic programming language for numerical computing". We employ Julia is to solve the Fokker-Planck equation by the stochastic approach with the use of the corresponding set of ordinary differential equations. We propose the method of parallelizing the algorithm with use of the distributed arrays. We test the speedup and efficiency of the given code with use of the cluster set at the Świerk Computing Centre and show that Julia is capable of achieving a good performance.

I. INTRODUCTION

CHOICE of the programming language for implementation of the mathematical model is quite a key factor for its future performance. Scientists routinely run simulations on millions of cores in distributed environments. Moreover, this choice is often influenced by the difficulty level of the language and time that has to spend on code production and its parallelization.

The main high-performance computing programming languages are statically compiled language such as Fortran, C, and C++, in conjunction with OpenMP/MPI. The reasons are their interoperability and efficiency regarding the ability to use all available compute resources while limiting memory usage. These languages are compiled off-line and have strict variable typing, allowing advanced optimizations of the code to be made by the compiler. Taking above into account one can think that the choice of a programming language is natural. However, it seems that writing HPC code is getting more complicated because today's projects often require a combination of messaging (MPI), threads (OpenMP), and accelerators (Cuda, OpenCL, OpenACC). This causes that creating an HPC code seems to be getting more difficult

instead of more straightforward. Of course, it is acceptable from the point of view of computer scientists or developers, but for the scientists from other fields more imperative is to get a relatively quick result to confirm/reject their theories or models.

The answer to that problems are the modern interpreted languages. In these languages, programs may be executed from source code form, by an interpreter. The advantages of this class are that they are often easier to implement in interpreters than in compilers, include platform independence, dynamic typing and dynamic scoping. The disadvantage is that they are usually slower than the compiled one. Examples of languages of this type are, e.g., Octave, Scilab, R, Mathematica, and Matlab. This category of languages is also known as dynamic languages or dynamically typed languages. In these programming languages, programmers write simple, high-level code without any mention of types like `int`, `float` or `double` that pervade statically typed languages such as C and Fortran. The overview of dynamical languages in comparison with the more traditional languages is presented in [1].

II. JULIA PROGRAMMING LANGUAGE

In this section we would like to introduce a Julia language, in which was implemented the algorithm presented in this paper. This will not be a tutorial, only some functions and issues useful in the parallel implementation of the proposed algorithm will be presented. For a full information we refer to [2].

Julia is a new programming language that is designed to address the problem of the low performance of dynamic languages [3]. In benchmarks reported by its authors, Julia performed within a factor of two of C on a set of standard basic tasks. Julia is advertised as a high-level, high-performance dynamic programming language for numerical computing. It provides a sophisticated compiler, distributed parallel execution, numerical accuracy, and an extensive mathematical function library. Julia's Base library, written mainly in Julia itself, also integrates mature, best-of-breed open source C and Fortran libraries for linear algebra, random

This work is supported by The Polish National Science Centre grant awarded by decision number DEC-2012/07/D/ST6/02488

number generation, signal processing, and string processing.

Julia is a free open source language that can be downloaded from the website [2]. The core of the Julia implementation is licensed under the MIT license. The language can be built as a shared library so that users can combine Julia with their own C/Fortran code or proprietary third-party libraries. The GitHub repository of Julia source code, packages, and documentation can be downloaded from website [4].

Users interact with Julia through a standard REPL (real-eval-print loop environment) such as Python, R, or MATLAB), by collecting commands in a .jl file, or by typing directly in a Jupyter (Julia, PYThon, R) notebook [5]. Julia syntax is "Matlab like" and allows uncomplicated expression of mathematical formulas. Julia uses just-in-time (JIT) compilation [6], [7] using the Low-Level-Virtual-Machine (LLVM) compiler framework [8]. JIT compilers attempt to compile at run-time by inferring information not explicitly stated by the programmer and use these inferences to optimize the machine code that is produced.

A. Parallel Computing in Julia

Julia has some built-in primitives for parallel computing at every level: vectorization (SIMD), multithreading, and distributed computing. At the lower level, Julia's parallel processing is based on the idea of remote references and remote calls. Remote call send a request to run a function on another processor, while remote reference create an object used to refer to objects stored on a particular processor. The parallel programming in Julia is quite intuitive. Starting the Julia in command line with `julia -p n` provides `n` worker processes on the local machine. It makes sense for `n` to be equal the number of CPU cores on the machine. In the script the additional `n` processes can be added by function `addprocs(n)` and removed by `rmprocs()`. Number of active processes can be listed by `workers()`.

Consider a parallel computation of the pi value by the Monte Carlo method. This calculation contains generating random numbers between 0 and 1 and ultimately calculating the ratio of the ones lying in inside the unit circle to those that don't.

```
addprocs(2) #add 2 Julia worker processes

function parallel_PI(n)
    in_circle = @parallel (+) for i in 1:n
        # work parallelizing
        x = rand()
        y = rand()
        Int((x^2 + y^2) < 1.0)
    end
    return (in_circle/n) * 4.0
end
```

The above function execution is:

```
parallel_PI(10000)
```

`@parallel` is for parallelizing loops. Julia offloads task to its worker processes that compute the desired output and send them back to the Julia master, where the reduction is performed. Arbitrary pieces of computation can be assigned to different worker processes through this one-sided communication model.

The algorithm presented in section III requires using large arrays for storing the pseudoparticles position in time of simulation. In such cases the distribution of the array between the workers seems to be a good solution. A distributed array is logically a single array, but its fragments are stored on several processors. Such approach allows making a matrix operation the same like with local arrays, making the parallelism almost invisible for the user. In some cases, it is possible to obtain useful parallelism just by changing a local array to a distributed array. Moreover, it makes possible to use an array of a size that wouldn't be possible to create in memory of one master process.

In Julia distributed arrays are implemented by the `DArray` type, which from version 0.4 has to be imported as the `DistributedArrays.jl` package from [9]. A `DArray` has an element type and dimensions just like a Julia array, but it also needs an additional property: the dimension along which data are distributed. Distributed array can be created in a following way:

```
julia>addprocs(4)
julia>@everywhere using DistributedArrays
julia>Tab1=drandn(8,8,4)
julia>Tab2=dfill(-1,8,8,4)
julia>Tab3=dzeros(8,8,4)
julia>Tab4=dzeros((8,8),workers()[1:4],[1,4])
```

In the above code, the four workers are started. Macro `@everywhere` allow precompiling the `DistributedArrays` on all processors. In the declaration of `Tab1`, `Tab2`, `Tab3` the distribution of this 8×8 arrays between four processors will be automatically picked. Random numbers will fill `Tab1`, `Tab2` will be filled with number `-1`, and `Tab3` with zeros. In the definition of the `Tab4` user can specify which processes to use, and how the data should be distributed. The second argument specifies that the array should be created on the first four workers. The third argument specifies on how many pieces chosen dimension should be divided into. In this example, the first dimension will not be divided, and the second dimension will be divided into 4 pieces. Therefore each local chunk will be of size $(8,2)$. The product of the distribution array must equal the number of workers.

This way of parallelization is quite convenient because when dividing data among a large number of processes, one often sees diminishing gains in performance. Placing `DArray` on a subset of processes allows numerous `DArray` computations to happen at once, with a higher ratio of work to communication on each process. Method `indexes` allow checking how the array is distributed. For example output of instruction

```
julia>Tab1.indexes
```

can be following

```
2x2x1 2x2x1 Array{Tuple{UnitRange{Int64},
  UnitRange{Int64},UnitRange{Int64}},3}:
[:, :, 1] =
(1:4,1:4,1:4) (1:4,5:8,1:4)
(5:8,1:4,1:4) (5:8,5:8,1:4)
```

We see that array is divided into four parts versus a number of rows and columns.

Other useful operations on distributed arrays are:

- `distribute(a::Array)` converts a local array to a distributed array,
- `localpart(a::DArray)` obtains the locally-stored portion of a DArray,
- `myindexes(a::DArray)` gives a tuple of the index ranges owned by the local process,
- `convert(Array, a::DArray)` brings all the data to the local processor.

When a DArray is created (usually on the master process), the returned DArray object stores information on how the array is distributed. When the DArray object on the master process is garbage collected, all participating workers are notified and localparts of the DArray freed on each worker. Since the size of the DArray object itself is small, a problem arises as `gc` on the master faces no memory pressure to collect the DArray immediately. This results in a delay of the memory being released on the participating workers. Therefore it is required to explicitly call `close(d::DArray)` after user code has finished working with the distributed array [9].

III. THE ALGORITHM FOR NUMERICAL SOLUTION OF THE FOKKER-PLANCK EQUATION

The Fokker-Planck equation (FPE) arises in a wide variety of natural science, including solid-state physics, quantum optics, chemical physics, theoretical biology and astrophysics. The FPE was first utilized by Fokker and Planck to describe the Brownian motion of particles e.g. [10]. In this paper we will use this equation to describe the transport of cosmic rays (CR) throughout the heliosphere, originating from outer space and reaching the Earth (e.g., [11]). The transport of CR particles is usually described by the Parker transport equation (PTE) [12]. The difficulty of the numerical solution of this type equations increases with the problem dimension. Reason is the instability of the numerical schemes like finite-differences (e.g. [13]) and finite-volume in the higher dimensions. In consequence, to ensure the scheme stability and convergence the density of numerical grid must be improved, increasing the computational complexity. To overcome this problem the stochastic methods can be applied. In that case the PTE should be rewritten in the form of the FPE (for details see, e.g. [14], [15]), as:

$$\frac{\partial \hat{f}}{\partial t} = \vec{\nabla} \cdot [\vec{\nabla} \cdot (K^T \hat{f})] - \vec{\nabla} \cdot [(\vec{\nabla} K^T + U) \cdot \hat{f}] + \frac{1}{3} \frac{\partial}{\partial R} [(\hat{f} R (\vec{\nabla} \cdot \vec{U})) - L \cdot \hat{f}]. \quad (1)$$

Where $\hat{f} = \hat{f}(\vec{r}, R, t)$ is an omnidirectional distribution function depending on spherical coordinates $\vec{r} = (r, \theta, \varphi)$, r - radial distance, θ - heliolatitudes, φ - heliolongitudes;

magnetic rigidity R and time t . $R = \frac{Pc}{q}$, where P is momentum, c speed of light, $q = Ze$, Z charge number of nucleus and e unit charge; \vec{U} is the solar wind velocity, K is the anisotropic diffusion tensor, K^T its transpose; L is the linear factor.

Applying the Ito stochastic integral we can bring the solution of the Eq. (1) to the solution of the set of stochastic ordinary differential equations (SDEs) being the exact equivalence of the FPE (e.g., [16]). Details of this procedure are given in ([14], [15] and references therein). Accordingly, the transport of CR in the 2D heliocentric coordinate system, in which $\vec{r} = (r, \theta)$, can be described by the following SDEs:

$$\begin{aligned} dr(t) &= \left(\frac{2}{r} K_{rr}^S + \frac{\partial K_{rr}^S}{\partial r} + \frac{ctg\theta}{r} K_{\theta r}^S + \frac{1}{r} \frac{\partial K_{\theta r}^S}{\partial \theta} + U + v_{d,r} \right) \cdot dt \\ &\quad + [B \cdot dW]_r \\ d\theta(t) &= \left(\frac{K_{r\theta}^S}{r^2} + \frac{1}{r} \frac{\partial K_{r\theta}^S}{\partial r} + \frac{1}{r^2} \frac{\partial K_{\theta\theta}^S}{\partial \theta} + \frac{ctg\theta}{r^2} K_{\theta\theta}^S + \frac{1}{r} v_{d,\theta} \right) \cdot dt \\ &\quad + [B \cdot dW]_\theta \\ dR(t) &= -\frac{R}{3} (\vec{\nabla} \cdot U) \cdot dt. \end{aligned} \quad (2)$$

In set of Eqs. (2) \vec{v}_d the drift velocity calculated as: $v_{d,i} = \frac{\partial K^A}{\partial x_j}$, where K^A is the antisymmetric part of the anisotropic diffusion tensor of the CR particles $K = K^S + K^A$ and $K^T = K^S - K^A$, containing the symmetric K^S and antisymmetric K^A parts given in [18]. The stochastic terms contain an element $d\vec{W}$ which is the increment of Wiener process guiding the stochastic motion of pseudoparticles in given dimension. B_{ij} , ($i, j = r, \theta$) is a matrix given in [15]. We discretize the set of Eqs. (2) with the unconditionally stable Euler-Maruyama [17] scheme. Nevertheless, the Eqs. (2) does not contain the linear factor L . Thus its solution is not synonymous with a solution of Eq. 1. In numerical realization we introduced weight W in which a linear factor L is taken into account according to formula:

$$W = \exp\left(-\int_0^t L(t) dt\right). \quad (3)$$

Consequently, the \hat{f} function value is expressed as a weighted average having the following form:

$$\begin{aligned} \hat{f}(\vec{r}, R) &= \frac{1}{N_f} \sum_{n=1}^{N_f} f_{LIS}(R) \cdot W = \\ &= \frac{1}{N_f} \sum_{n=1}^{N_f} f_{LIS}(R) \cdot \exp\left(-\sum_{m=1}^M L_{,m} \cdot \Delta t\right). \end{aligned} \quad (4)$$

$L = -\frac{2}{3} \vec{\nabla} \cdot U$ is the linear factor visible in Eq. 1, N is the total number of simulated pseudoparticles, N_f is the number of pseudoparticles reaching the position \vec{r} and M is the number of time steps. Function $f_{LIS}(R)$ denotes the \hat{f} value at the boundary.

During the simulation pseudoparticles are initialized at the region (heliosphere) boundary with the initial rigidity drawn by the rejection sampling algorithm from $f_{LIS}(R)$ distribution; then their trajectories are traced in conjunction with changes of their rigidity R . The position and rigidity of each pseudoparticle in every time step must be stored to find the value of the distribution function $\hat{f}(\vec{r}, R)$ in each (required) point of the region. The pseudoparticle motion is

terminated when it reaches the inner/outer boundary with respect to the radial distance or when the time for simulation finishes. As far as the probability that statistically enough number of pseudoparticles will reach the single point is near to zero, we use the bins instead of the points. Thus, to find the numerical solution of FPE it is necessary to apply the binning procedure, i.e., discretized the 3D domain over all spatial variables: (r, θ) and R . Then for each binning unit $[r \pm \Delta r] \times [\theta \pm \Delta \theta] \times [R \pm \Delta R]$, we integrate the trajectories of pseudoparticles traveling through considered bin according to Eq. 4. The binning procedure is the most time consuming from the computational point of view.

IV. JULIA PARALLEL CODE FOR THE SOLUTION OF FPE

The code for the numerical solution of the set of Eqs. (2) was realized in Julia v 0.6.2 [2]. The stochastic method solution of FPE is quite easy to parallelize versus the number of simulated pseudoparticles, which can be simulated independently. To do this large arrays are required for storing the pseudoparticles position, rigidity and weight in subsequent time steps. A natural way to obtain parallelism is to distribute arrays between many processors. This approach combines the memory resources of multiple machines, allowing to create and operate on arrays that would be too large to fit on one machine. Each processor operates on its own part of the array, making possible a simple and quick distribution of task among machines. A distributed array is logically a single array, but its fragments are stored on several processors. Such approach allows making a matrix operation the same like with local arrays, making the parallelism almost invisible for the user. This way was written the parallel program solving the Eq. 1 by the method described in detail in Section III.

The proposed construction of Julia code solving FPE by the stochastic approach described in section III is given in Algorithm 1.

Command in first line calls the required number of CPUs. In lines 5-9 the tables storing the pseudoparticles position and rigidity are distributed among the processes via the `DistributedArrays` package. The distribution is done versus the second dimension, i.e. a single processor covers array of size $n \times (m/WN)$. The simulation of pseudoparticles motion is done by the function `SEQ()` given in lines 12-36. The pseudoparticle initial position and rigidity is set in line 16-18; its position is changed accordingly to the equation 2 (lines 15-19) including the Wiener process defined in line 20-21. The initial weight is set in line 19, while its change is set in line 31. The simulation is performed until all pseudoparticles meet the termination conditions (line 33). In lines 40-52 is defined the function `PRL()` which runs the function `SEQ()` in parallel on different workers owned by the distributed arrays. The `@spawnat` macro evaluates the expression in the second argument on the process specified by the first argument. Function `pmap()` transform collection `out` by applying `fetch` to each element using available workers and tasks. The actual launch of function `PRL()` is done in line 55.

Algorithm 1 Draft of Julia parallel code for FPE solution

```

1  addprocs(number_of_processors)
2  n=number_of_time_steps;
3  m=number_of_pseudoparticles;
4  WN=length(workers());
5  @everywhere using DistributedArrays
6  r=dzeros((n,m),workers()[1:WN],[1,WN])
7  T=dzeros((n,m),workers()[1:WN],[1,WN])
8  R=dzeros((n,m),workers()[1:WN],[1,WN])
9  W=dzeros((n,m),workers()[1:WN],[1,WN])
10 #####
11
12 @everywhere function SEQ(n,m,r,T,R,W)
13   for j = 1 : m
14     #defining pseudoparticles
15     #initial characteristics in t=0;
16     r[1,j]=...;
17     T[1,j]=...;
18     R[1,j]=...;
19     W[1,j]=1;
20     dWr=Wiener(n);#generating the Wiener
21     dWt=Wiener(n);#processes
22     for i = 1 : n-1
23       #calculation of the dr, dT,
24       #dR, dW according to Eqs.2
25       ...
26       #calculation of new pseudopart.
27       #characteristics
28       r[i+1,j]=r[i,j]+dr[i,j];
29       T[i+1,j]=T[i,j]+dT[i,j];
30       R[i+1,j]=R[i,j]+dR[i,j];
31       W[i+1,j]=W[i,j]*exp(-Lf*dt);
32       boundary_verification;
33       termination_verification;
34     end
35   end
36 end
37
38 #####
39
40 function PRL(n,m,r,T,R,W)
41   P=length(procs(r))
42   Nlocal=[size((r.indexes)[w][1],1)
43           for w=1:P]
44   Mlocal=[size((r.indexes)[w][2],1)
45           for w=1:P]
46   out=[(@spawnat(procs(r))[w]
47         SEQ(Nlocal[w],Mlocal[w],
48            localpart(r),localpart(T),
49            localpart(R),localpart(W))
50         for w=1:P]
51   pmap(fetch,out)
52 end
53 #####
54
55 @time PRL(n,m,r,T,R,W) #actual run
56 @time PRLBin(n,m,r,r_bin,T,T_bin,R,R_bin,W)

```

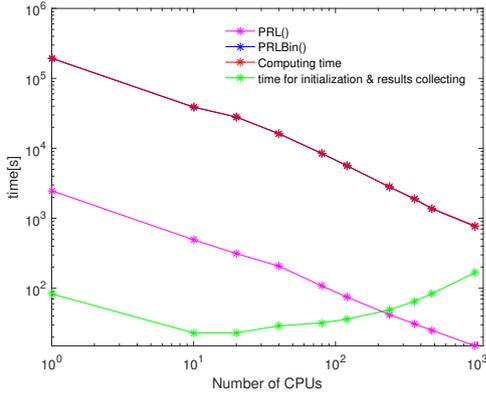


Fig. 1. The time of the FPE solution in Julia ver 0.6.2 on CIŚ cluster. Figure in log-log scale.

To obtain the value of the distribution function we have to run the function `PRLBin()` which search for pseudoparticles falling into the bins $[r \pm \Delta r] \times [\theta \pm \Delta \theta] \times [R \pm \Delta R]$ and apply the Eq. 4 to get the $f(\vec{r}, R)$ value in each bin. The macro `@time` measures the performance of the function run by returning the calculation time and amount of allocated memory.

The launching, management, and networking of Julia processes into a cluster can be done via `ClusterManager.jl` package [19]. It supports different job queue systems commonly used on computer clusters as Slurm, Sun Grid Engine, and PBS. However, this package doesn't support the Torque system installed on CIŚ cluster used to perform simulations presented in this paper. In such case, the distribution can be done directly via the `machinofile`. The sample `.PBS` file that is send to the queue can have a form:

```
#!/bin/bash
#PBS -N task_name
#PBS -l nodes=N :ppn=P
#PBS -q queue_name
cd $PBS_O_WORKDIR
julia --machinofile $PBS_NODEFILE
/mnt/home/user_catalogue/task_code.jl
```

In above-presented script the number of required nodes is equal to N , and number of CPU's per node to P . Thus the $N * P$ workers will be allocated to the job. The names of all nodes the job has allocated, with an entry for every CPU will be saved to the `nodefile`. Thus Julia will read the number of workers from that file, so calling the `adprocs()` in the first line of Algorithm 1 should be omitted.

A. Julia performance

The presented algorithm complexity is $O(n)$ and depends on the number of pseudoparticles and number of time steps. To obtain the reliable results and achieve the satisfactory statistics

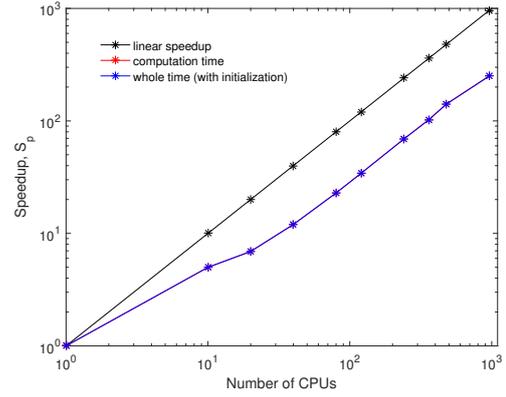


Fig. 2. The calculated speed-up of the FPE solution in ver Julia 0.6.2 on CIŚ cluster. The black line shows the case of ideal speed-up. Figure in log-log scale.

in each 2D heliosphere bin we have to run at least 2 millions of pseudoparticles from the heliosphere boundary. In this paper we do not focus on the results of the physical model, interested reader we refer to the following papers [14], [15]. Here we would like to test the performance of above presented parallel Julia code on the HPC cluster. We have employed the CIŚ machine with characteristic given in Table I.

We have run series of simulations changing the number of CPUs in the range from 1 up to 960. Used model was simplified, as far as it should be run on the one CPU. We assumed the simulation time to be equal to $t = 225$ days with a time step $\Delta t = 1$ hour and a number of simulated pseudoparticles $m = 7200$.

We have analyzed the results accordingly to Amdahl's model [20] that assumes the problem size does not change with the number of CPUs and wants to solve a fixed-size problem as quickly as possible. The model is also known as speedup, which can be defined as the maximum expected improvement to an overall system when only part of the system, is improved. We have used for evaluating the performance of the parallel code the parallel runtime, the speedup S_p calculated as:

$$S_p = \frac{T_s}{T_p}, \quad (5)$$

where T_s is sequential runtime using one CPU, and T_p is runtime using p -number of CPUs. We have also estimated the efficiency E_p as:

$$E_p = \frac{T_s}{p * T_p}. \quad (6)$$

We have run calculations assuming the job distribution over the 1, 10, 20, 40, 80, 120, 240, 360, 480, and 960 CPUs. The number of used CPU has a direct impact on number of pseudoparticles simulated on single CPU, i.e. 7200, 720, 360, 180, 90, 60, 30, 20, 15 and 7, respectively. The results of parallel runtime for the computing the whole simulation, particular functions and the time required for initialization and data collecting present Fig. 1. The computing time includes

TABLE I
THE USED CIŚ MACHINE CHARACTERISTICS.

Feature	Specification
Model	Intel S2600TP
CPU	Intel Xeon(R) CPU E5-2680v2
CPU frequency	2.8Hz, up to 3.6GHz in turbo mode
CPUs per node	40
RAM per node	128 GB, DDR3
Interconnect	1 x Ethernet (1Gbit/sec per port)

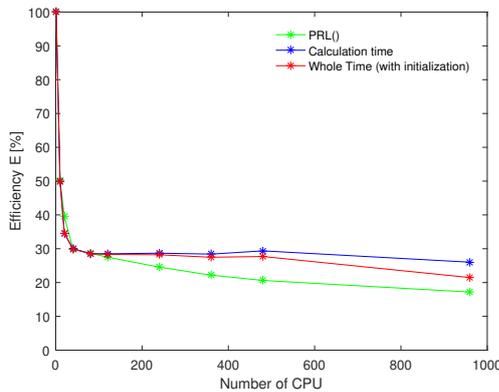


Fig. 3. The efficiency of the FPE solution in ver Julia 0.6.2 on CIŚ cluster.

both the `PRL()` and `PRLBin()` functions parallel execution times. The binning procedure has the most substantial contribution to computing time. We see that the execution time decreases with CPUs number quite steady, simultaneously, the time for initialization and collecting the data increases, because more communication between the nodes is required. The Fig. 2 presents the speedup. We can see that the presented method of parallelization gives a good sublinear speedup. The speedup is behaving very stable and follows a straight line starting from 1 node (40 CPU). Up to one node, the line declines from a straight line. The reason might be that the other tasks overloaded the free CPUs consuming large memory. The corresponding efficiency rate presents the Fig. 3. The efficiency curve shows a negative gradient as the efficiency reduces with an increased number of processors. It stabilizes starting from 80 CPUs up to 480 CPUs at level of 28%, then decreases up to 21%. This metric measures the effectiveness of parallel algorithm concerning the computation time. These results show also that the application of the parallelization based on the distributed arrays is quite efficient in Julia.

V. SUMMARY

We have implemented the proposed method in the new high-level Julia programming language. Parallelization of the code is based on decomposing the problem into a subset of independent tasks versus the number of simulated pseudoparticles. We recognized Julia as a very suitable intuitive

tool for parallel implementation. We have analyzed the HPC performance of Julia as the speedup and efficiency with the use of the CIŚ cluster. We can conclude that the performance of Julia code utilizing the distributed arrays is quite good. Moreover, application of the Julia built-in parallelization methods based on remote references and remote calls does not require from the user much effort or additional work/knowledge on serialization and message passing between workers. These features allow recommending Julia in HPC calculations in the cases when results should be archived relatively quickly and a small amount of time can be taken for the code parallelization.

ACKNOWLEDGMENT

Calculations were performed at the Świerk Computing Centre being a part of the National Centre for Nuclear Research.

REFERENCES

- [1] Rei, L., Carvalho, S., Alves, M., Brito, J., A look at dynamic languages, *Tech. report*, 2007, Faculty of Engineering University of Porto.
- [2] The Julia Language <https://julialang.org/>
- [3] Bezanson J. et. al., Julia: A Fresh Approach to Numerical Computing, *SIAM Review*, vol. 59, 1 (2017) 65-98
- [4] GitHub <https://github.com/JuliaLang/julia>
- [5] The Jupyter Project, <http://jupyter.org/>
- [6] J. Bezanson, Abstraction in Technical Computing, Ph.D. thesis, Massachusetts Institute of Technology, MA, 2015.
- [7] Bezanson J., S. Karpinski, V. B. Shah, and A. Edelman, Julia: A Fast Dynamic Language for Technical Computing, preprint, arXiv:1209.5145 [cs.PL], 2012.
- [8] Lattner C., Adve V., LLVM: A compilation framework for lifelong program analysis & transformation. In: *Proceedings of the international symposium on Code generation and optimization: feedback-directed and runtime optimization*. IEEE Computer Society, 2004, 75
- [9] Distributed Arrays Packadge <https://github.com/JuliaParallel/DistributedArrays.jl>
- [10] Risken, H., *The Fokker-Planck Equation Method of Solution and Applications*, SpringerVerlag, Berlin, Heidelberg, 1989
- [11] Moraal H., Cosmic-Ray Modulation Equations, *Space Science Reviews*, 2013, vol. 176, 299-319 doi:10.1007/s11214-011-9819-3
- [12] Parker E. The passage of energetic charged particles through interplanetary space, *Planetary and Space Science*, 1965, vol. 13, 9-49
- [13] Wawrzynczak A., Alania M.V., Numerical Solution of the Time and Rigidity Dependent Three Dimensional Second Order Partial Differential Equation, *Lecture Notes in Computer Science*, 2010, vol. 6067, pp. 105-114, doi: 10.1007/978-3-642-14390-8_12
- [14] Wawrzynczak A., Modzelewska R and Gil A, Stochastic approach to the numerical solution of the non-stationary Parker's transport equation, *Journal of Physics: Conference Series*, 2015, vol. 574, 012078, doi:10.1088/1742-6596/574/1/012078
- [15] Wawrzynczak A., Modzelewska R and Gil A, The algorithms for forward and backward solution of the Fokker-Planck equation in the heliospheric transport of cosmic rays, *Lecture Notes in Computer Science*, 2018, vol. 10777, 14-23, doi:10.1007/978-3-319-78024-5-2
- [16] Gardiner C.W., *Handbook of stochastic methods. For physics, chemistry and the natural sciences*, Springer Series in Synergetics, 2009
- [17] Kloeden P E, Platen E, Schurz H., *Numerical solution of SDE through computer experiments*, Springer-Verlag Berlin Heidelberg, 1992
- [18] Alania M. V., Stochastic Variations of Galactic Cosmic Rays, *Acta Physica Pol. B*, 2002, vol. 33(4), 1149-1166
- [19] Cluster Manager Packadge <https://github.com/JuliaParallel/ClusterManagers.jl>
- [20] Amdahl, G.M., Validity of the single-processor approach to achieving large scale computing capabilities, in: Proc. Am. Federation of Information Processing Societies Conf., AFIPS Press, (1967) 483-485.

11th International Symposium on Multimedia Applications and Processing

SFTWARE Engineering Department, Faculty of Automation, Computers and Electronics, University of Craiova, Romania “Multimedia Applications Development” Research Centre

BACKGROUND AND GOALS

Multimedia information has become ubiquitous on the web, creating new challenges for indexing, access, search and retrieval. Recent advances in pervasive computers, networks, telecommunications, and information technology, along with the proliferation of multimedia mobile devices—such as laptops, iPods, personal digital assistants (PDA), and cellular telephones—have stimulated the development of intelligent pervasive multimedia applications. These key technologies are creating a multimedia revolution that will have significant impact across a wide spectrum of consumer, business, healthcare, educational and governmental domains. Yet many challenges remain, especially when it comes to efficiently indexing, mining, querying, searching, retrieving, displaying and interacting with multimedia data.

The Multimedia—Processing and Applications 2018 (MMAAP 2018) Symposium addresses several themes related to theory and practice within multimedia domain. The enormous interest in multimedia from many activity areas (medicine, entertainment, education) led researchers and industry to make a continuous effort to create new, innovative multimedia algorithms and applications.

As a result the conference goal is to bring together researchers, engineers, developers and practitioners in order to communicate their newest and original contributions. The key objective of the MMAAP conference is to gather results from academia and industry partners working in all subfields of multimedia: content design, development, authoring and evaluation, systems/tools oriented research and development. We are also interested in looking at service architectures, protocols, and standards for multimedia communications—including middleware—along with the related security issues, such as secure multimedia information sharing. Finally, we encourage submissions describing work on novel applications that exploit the unique set of advantages offered by multimedia computing techniques, including home-networked entertainment and games. However, innovative contributions that don't exactly fit into these areas will also be considered because they might be of benefit to conference attendees.

CALL FOR PAPERS

MMAAP 2018 is a major forum for researchers and practitioners from academia, industry, and government to present, discuss, and exchange ideas that address real-world problems with real-world solutions.

The MMAAP 2018 Symposium welcomes submissions of original papers concerning all aspects of multimedia domain ranging from concepts and theoretical developments to advanced technologies and innovative applications. MMAAP 2018 invites original previously unpublished contributions that are not submitted concurrently to a journal or another conference. Papers acceptance and publication will be judged based on their relevance to the symposium theme, clarity of presentation, originality and accuracy of results and proposed solutions.

TOPICS

- Audio, Image and Video Processing
- Animation, Virtual Reality, 3D and Stereo Imaging
- Big Data Science and Multimedia Systems
- Cloud Computing and Multimedia Applications
- Machine Learning, Information Retrieval in Multimedia Applications
- Data Mining, Warehousing and Knowledge Extraction
- Multimedia File Systems and Databases: Indexing, Recognition and Retrieval
- Multimedia in Internet and Web Based Systems
- E-Learning, E-Commerce and E-Society Applications
- Human Computer Interaction and Interfaces in Multimedia Applications
- Multimedia in Medical Applications and Computational biology
- Entertainment, Personalized Systems and Games
- Security in Multimedia Applications: Authentication and Watermarking
- Distributed Multimedia Systems
- Network and Operating System Support for Multimedia
- Mobile Network Architecture and Fuzzy Logic Systems
- Intelligent Multimedia Network Applications
- Future Trends in Computing System Technologies and Applications
- Trends in Processing Multimedia Information
- Multimedia Ontology and Perception for Multimedia Users

BEST PAPER AWARD

A best paper award will be made for work of high quality presented at the MMAP Symposium. The technical committee in conjunction with the organizing/steering committee will decide on the qualifying papers. Award comprises a certificate for the authors and will be announced on time of conference.

STEERING COMMITTEE

- **Amy Neustein**, Boston University, USA, Editor of Speech Technology
- **Lakhmi C. Jain**, University of South Australia and University of Canberra, Australia
- **Zurada, Jacek**, University of Louisville, United States
- **Ioannis Pitas**, University of Thessaloniki, Greece
- **Costin Badica**, University of Craiova, Romania
- **Borko Furht**, Florida Atlantic University, USA
- **Harald Kosch**, University of Passau, Germany
- **Vladimir Uskov**, Bradley University, USA
- **Thomas M. Deserno**, Aachen University, Germany

HONORARY CHAIR

- **Dumitru Dan Burdescu**, University of Craiova, Romania

GENERAL CO-CHAIRS

- **Adriana Schiopoiu Burlea**, University of Craiova, Romania
- **Marius Brezovan**, University of Craiova, Romania

PUBLICITY CHAIR

- **Amelia Badica**, University of Craiova, Romania
- **Milan Simic**, RMIT University, School of Engineering, Australia

ORGANIZING

- **Dumitru Dan Burdescu**, University of Craiova, Romania
- **Costin Badica**, University of Craiova, Romania
- **Marius Brezovan**, University of Craiova, Romania
- **Adriana Schiopoiu Burlea**, University of Craiova, Romania
- **Liana Stanescu**, University of Craiova, Romania
- **Cristian Marian Mihaescu**, University of Craiova, Romania

PROGRAM COMMITTEE

- **Azevedo, Ana**, CEOS.PP-ISCAP/IPP, Portugal
- **Badica, Amelia**, University of Craiova, Romania
- **Burlea Schiopoiu, Adriana**, University of Craiova, Romania
- **Cano, Alberto**, Virginia Commonwealth University
- **Cordeiro, Jose**, EST Setúbal/I.P.S.
- **Cretu, Vladimir**, Politehnica University of Timisoara, Romania
- **Debono, Carl James**, University of Malta, Malta

- **Fabijańska, Anna**, Lodz University of Technology, Poland - Institute of Applied Computer Science, Poland
- **Fomichov, Vladimir**, National Research University Higher School of Economics, Moscow, Russia., Russia
- **Giurca, Adrian**, Brandenburg University of Technology, Germany
- **Grosu, Daniel**, Wayne State University, United States
- **Kabranov, Ognian**, Cisco Systems, United States
- **Keswani, Dr. Bright**, Suresh Gyan Vihar University, Mahal, Jagatpura, Jaipur
- **Korzhih, Valery**, State University of Telecommunications, Russia
- **Kostagiolas, Petros**, School of Information Science and Informatics, Ionian University
- **Kotenko, Igor**, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science, Russia
- **Logofatu, Bogdan**, University of Bucharest, Romania
- **Mangioni, Giuseppe**, DIEEI - University of Catania, Italy
- **Marghitu, Daniela**, Auburn University
- **Mihaescu, Cristian**, University of Craiova, Reunion
- **Mocanu, Mihai**, University of Craiova, Romania
- **Murawski, Krzysztof**, Faculty of Cybernetics, Military University of Technology, Poland
- **MURAWSKI, Krzysztof**, Military University of Technology, Poland
- **Ohzeki, Kazuo**, Professor Emeritus at Shibaura Institute of Technology, Japan
- **Pohl, Daniel**, Intel, Germany
- **Popescu, Dan**, CSIRO, Sydney, Australia, Australia
- **Popescu, Daniela E.**, Integrated IT Management Service, University of Oradea
- **Querini, Marco**, Department of Civil Engineering and Computer Science Engineering
- **Radulescu, Florin**, University "Politehnica" of Bucharest
- **RUTKAUSKIENE, Danguole**, Kaunas University of Technology
- **Salem, Abdel-Badeeh M.**, Ain Shams University, Egypt
- **Sari, Riri Fitri**, University of Indonesia, Indonesia
- **Sousa Pinto, Agostinho**, Instituto Politécnico do Porto
- **Stoicu-Tivadar, Vasile**, University Politehnica Timisoara
- **Trausan-Matu, Stefan**, Politehnica University of Bucharest, Romania
- **Trzcielinski, Stefan**, Poznan University of Technology, Poland
- **Tsahrintzis, George**, University of Piraeus, Greece
- **Tudoroiu, Nicolae**, John Abbott College, Canada
- **Vega-Rodríguez, Miguel A.**, University of Extremadura, Spain
- **Virvou, Maria**, University of Piraeus, Greece
- **Watanabe, Toyohide**, University of Nagoya
- **Woźniak, Marcin**, Institute of Mathematics, Silesian University of Technology, Poland

Developing keyword spotting method for the Polish language

Lukasz Laszko

Cybernetics Faculty,
Military University of
Technology,
ul. Gen. W. Urbanowicza 2,
00-908 Warsaw, Poland
Email:
lukasz.laszko@wat.edu.pl

Abstract—The paper presents the application of unsupervised method to word detection in recorded speech for the spoken Polish language. The method utilizes similarity measure between analyzed speech and a pattern synthesized from pure text. Dynamic time warping algorithm is applied for time alignment and the resulting alignment path defines an input to the classifier. The classification process involves calculation of cost function and extraction of the projected sequence of Human-Factor Cepstral Coefficients, both of which are compared with the threshold values. The results obtained after application of the method to the CLARIN-PL Mobile Corpus are encouraging to develop this method for the Polish language.

I. INTRODUCTION

THE hankering for good and robust method of automatic speech recognition or word spotting for the Polish language has been observed for years in Polish scientific, as well as business field. Recently much effort has been put to appropriate language modeling [1], considering the nature of the Polish spoken language. These studies reveal the complexity of the modeling process for general purposes and indicate particularly difficult attributes of the language to model, such as its inflection, non-positionality and frequent occurrence of short words. Deep insight into speech signal shows also that the existence of high-frequency, low-energy consonants like fricatives and plosives, restricts the adopting of widely used methods and tools good for the English language [3]. Although to simple daily tasks one could employ with success the grammar-based ASR's [2], which firstly require primitive ontological relations to be built for a class of sentences in the given field. Either HMM or various classes of neural networks, built on specific acoustic features, are the most common models used in this area.

Considering the evolution of speech recognition and speech processing tools available for the Polish language the trend to exploit open-source technologies is observed. One example is SARMATA [4], the aboriginal Polish ASR

This work was supported by Cybernetics Faculty of the Military University of Technology, under the grant no. RMN/813/2016

system, which has recently (version 2.0) being under departing from its own engine to Kaldi toolkit. The system in its pre-2.0 versions was able to be used in industry, recognizing up to 1000 learned words. The new version is very likely to be much more versatile, because of using available in the toolkit large number of possible to use speech models and techniques of speech processing, as well as massive GPU processing implemented in the toolkit.

Contemporarily, the most-growing Polish set of tools, as it seems to be, is provided by CLARIN-PL. This is actually much more than the set of software, but it is seen as a speech platform for processing, visualizing and depositing language data [5]. This platform provides cloud-based research infrastructure (type B) with corpora, tools (via web services¹) and metadata. It also enables users to make available their own products like tools or corpora².

Nor the reader can miss the *de facto standard* of Google SpeechRecognizer, which engine is integrated to most Android mobile devices used today, and has a support for 119 languages including the Polish language. Moreover its API is freely available to developers of Android applications.

In this field the author propose the adoption of keyword spotting method (abbr. KWS) introduced in [6] for the spoken English language, to the Polish language. The method is designed to search for specific words only and does not analyze the structure of speech at higher levels than the acoustic features, i.e. the language or the grammar. There is also no supervised model training step, apart from that one do need to assess a few threshold values. Although applied for the English language, the method gives relatively high detection rate, about 80%.

II. PROBLEM STATEMENT

A. Method background

The precise description of the method as well as alternatives could be found in [6]. In the nutshell, this method is searching through a speech medium (database) fragment by

¹ For tools availability, see: <https://clarin-pl.eu/en/services/>

² E.g. *Acoustic Data Building Toolset*, about 29 hours (17 GB) corpus of annotated Polish speech, together with software.

fragment and comparing the description of each fragment with the same class of description of a search pattern. The description is understood as a sequence of acoustic features. The comparison is done in the similarity space, which contains implicit information about correlation between the two descriptions. The strength of similarity is measured by applying Dynamic Time Warping (abbr. DTW) algorithm and extracting the best projection path of the search pattern description to the description of the analyzed fragment of speech, which is called in the method the alignment path. DTW fulfills therefore two important tasks: (1) time alignment between search pattern and analyzed fragment, (2) cost counting, which provides values to the classification process (see Fig 1).

In the classification stage the calculation of cost function and sequence search according to the extracted aligning path are done and compared with the threshold values. This provides the decision on the class of analyzed fragment of speech as match or no-match. Then upon the results of assessing values of cost function and found sequences, the quality of the matches is calculated, which provides the numerical control of matches.

Next stages depend on the application and could involve, but are not limited to the verification by listening or thorough search of the area pointed by best matches.

It is worth noting that one input to the model from Fig 1, comes from the Text-to-Speech generation. This is the valuable attribute of the presented method, which according to [6], makes the method versatile.

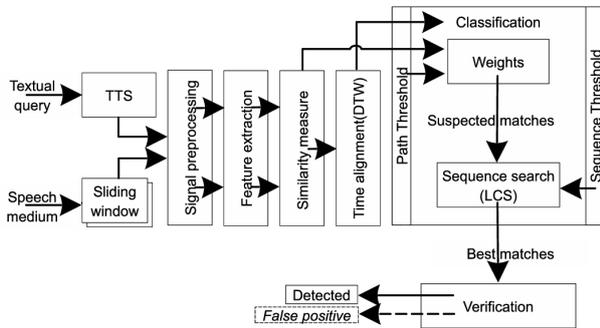


Fig 1. Overview of the unsupervised detection method

B. Mathematical model of speech description

In these research Human-Factor Cepstral Coefficients (abbr. HFCC) have been chosen as the description of speech signal. HFCCs are computed according to the following algorithm:

- 1) given signal S has been windowed by Hamming window resulting in N segments, $S_1 \dots S_N$;
- 2) each segment has been processed by short-time Fourier transform (abbr. STFT) with length of 64 ms and a fixed step size of 5 - 10 ms;
- 3) then the triangular filter bank has been developed with 40 equally spaced mel-scale center frequencies f_i ,

$i=1, \dots, 40$ with bands controlled by the measure called Equivalent Rectangular Bandwidth (abbr. ERB):

$$ERB(f) = 6.23 f^2 + 93.39 f + 28.52 \text{ Hz}; \quad (1)$$

where f states for filter center frequency, expressed in kHz.

4) next, the filtering has been done, by multiplication of each STFT segment with magnitude spectrum of bands for HFCC;

5) finally, the result has been decorrelated using Discrete Cosinus Transform (abbr. DCT), keeping only 15 the most decorrelated coefficients.

C. Measuring similarity of two signals

Let the matrix $D_{A,R}$, where A stands for analyzed voice feature vector and R stands for reference pattern feature vector, hold the information on similarity between A and R . Then the individual element $d(a,r)$ of the matrix, where a,r stand for specific element of vector A and vector R respectively, is given by inner product:

$$d(a,r) = \frac{\langle A_a, R_r \rangle}{\|A_a\| \|R_r\|} \quad (2)$$

D. Applying DTW

Let the $C_{A,R}$ be the accumulator matrix of size D . Then the accumulation in each element $c(a,r)$ holds the value of lowest transition cost to this element from its neighbors, including the cost of the lowest transition to the neighbors from their consequent neighbors until the starting element $c(1,1)$. The computation is given by the recursion:

$$c(a+1, r+1) = d(a+1, r+1) + \min \begin{cases} c(a-1, r) \\ c(a, r) \\ c(a, r-1) \end{cases} \quad (3)$$

where: $a, r \geq 1$ and $c(1,1) = d(1,1)$.

The stage of applying DTW gives the calculation of optimal aligning P of analyzed voice description and the reference pattern description. P is created based on the accumulator elements traceback, starting from its last element $c(N_A, N_R)$ and ending in $c(1,1)$ recursively, by searching across all allowable predecessors to each element. Because C has been built of costs of the lowest transitions, the actual calculation of the path is based on choosing the next element from the closest elements with minimal cost value.

E. Classification and quality measure

P and D hold then the full information on the similarity strength between analyzed fragment and referenced pattern. Upon this v is computed based on referring costs

of matrix $D_{A,R}$, where A, R are taken from P . Then v is equated to path threshold T_P , producing suspected matches M . To this result the Longest Common Subsequence (abbr. LCS) algorithm is applied to reject the least valuable sequences according to sequence threshold T_S . For all accepted results the quality measure is computed according to (4).

$$Q_M = \frac{v(M)}{LCS(M, T_S)} 100 \quad (4)$$

F. Variables

The method has many variables which values decide on the usability of the method. There is a need for: calculation of the width of analysis window; HFCC computation parameters which are used in point B.1, deciding on the feature space dimension discussed in point 5; P -specific calculations in DTW algorithm (direction variation, analyzed area in D , etc); threshold values: T_P and T_S which decide on the resultant matches, as well as the minimal quality value satisfactory for specific applications.

III. EXPERIMENTS

A. Research material

The experiments have been conducted on CLARIN-PL Mobile Corpus (EMU) [8]³. This is a Polish speech corpus of read speech recorded over a phone. It contains 554 sessions of many speakers reading a few dozen different sentences. Each recorded speech is annotated. Total corpus length is about 13 hours (12 GB uncompressed). Sound quality is at medium level (16 kHz, 32 bits/sample, mono) stored in WAV containers.

The queries have been generated using Google Text-to-Speech engine, available via Google Translate, based on a textual input.

B. Procedure

According to Fig 1, the experiments started by preparing queries and signal preprocessing. Then in accordance with point II.B, HFCC features were computed for the query as well as for the analyzed fragment. Fragments lengths were in these experiments 1.5 times longer in time than the queries. The last analyzed fragment was complemented with zeros.

Sliding windows were produced with fixed step size without overlapping of neighboring windows. The overlapping was included at the stage of computing HFCCs.

Then the D matrix was computed, before applying it DTW. Based on the results of DTW the classification of the

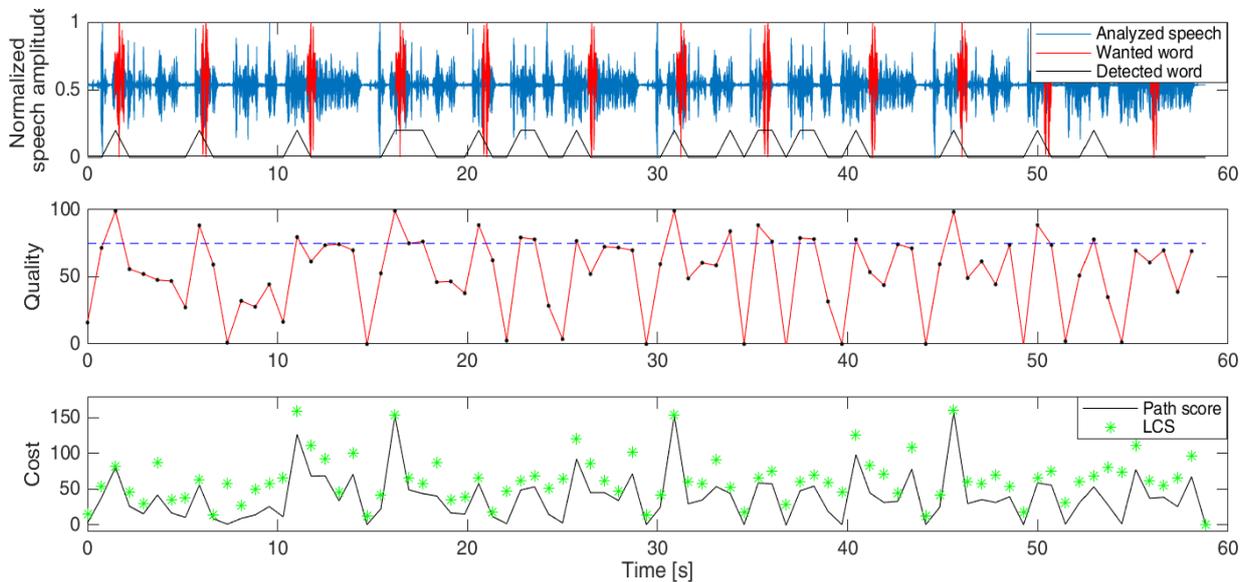


Fig 2. Results of word spotting on sentence 1 of session 1 of the CLARIN-PL Mobile Corpus. The chosen word ‘senator’ occurs 12 times in the sentence (in different inflections), which is marked in the upper chart. The upper chart has also markers placed around the bottom axis, to indicate the best matches obtained by the discussed method. Medium chart presents quality of detection, with the satisfying cutoff level of 75%. Bottom chart presents costs computed during classification stage.

³ For corpus imprint see: <https://clarin-pl.eu/dspace/handle/11321/237>

fragment was done thus obtaining the list of potential matches of sequences allegedly detected in the fragment.

Finally for accepted matches, the detection quality was assessed based on (4).

The procedure was repeated two times for the same sentence with the change of reference pattern. For method verification, the synthesized query was exchange for the excerpt of the same material with the same content.

C. Results and discussion

Overall results have been presented in Table 1. These results concern whole research material, which includes chosen sessions from 554 available sessions in the speech corpus, without distinction to the gender of speaker. Unfortunately only one TTS system has been used in the research (female voice of medium quality), which probably caused understating of the percentage of detected words.

High word detection rate has been observed. Concerning real speech pattern results, more values have been obtained over the presented mean value (negative skewness). Although false detection rate also maintained at rather high level, these results do not seem to correlate (correlation coefficient, CC equals: -0,2).

Referring to TTS results, positive impression is given, not only by high detection rate, but also by the maximal value for detection. This means that synthetic speech has perfectly been aligned to the real (unknown) speech in some experiments. Unfortunately this seems to correlate with false detection rate (CC equals: 0,6).

Fig 2 presents the exact results of an exemplary analysis. The analyzed speech has been manually replicated four times for the sake of observing method correctness for the relating fragments of speech. As presented in the upper chart, the best matches indicated by markers placed around the bottom axis, are in the area close to the place where the wanted word is spoken. These markers show eleven areas out of twelve occurrences of the word in the analyzed speech. Four markers point faulty and four other markers are redundant, because of pointing the area being already pointed.

The presented markers come from the quality assessment of the corresponding matches, which is presented entirely in the medium chart.

The bottom chart of this figure shows the costs computed during classification stage. Path score plot presents the v vectors of the corresponding path P , while the green stars present chosen subsequences extracted from M .

TABLE 1. OVERALL RESULTS BY SPEECH SOURCE

	Detected words	No detection	False detection
Real speech	82,92%	17,08%	56%
<i>min</i>	37,5%	0%	26,7%
<i>mode</i>	91,7%	0%	4%
<i>max</i>	100%	62,5%	75,7%
<i>Skewness</i>	-1,4	1,4	-0,6
TTS	74,17%	25,83%	41,76%
<i>min</i>	50%	0%	0%
<i>mode</i>	50%	50%	50%
<i>max</i>	100%	50%	66,7%
<i>Skewness</i>	0	0	-1,1
<i>Standard Error of the Mean: ~6%</i>			

Chosen fragments of the analyzed speech during the searching for the word ‘senator’ have been presented in Fig 3. Times in the titles of each charts indicate real time range related to the speech presented in the upper chart of Fig 2. Presented steps 3, 9 and 16 show the best matches for the word ‘senator’ found in analyzed speech. Time steps of the best matches are not presented for the sake of readability. Although this outline shows that length of matches are different (i.e. the red stripes vary in length).

Steps 2, 8, 10 and 15 show larger sections of the fragments with silence in speech. Normally DTW algorithm includes this in the alignment path, causing matches that not necessarily carrier important information.

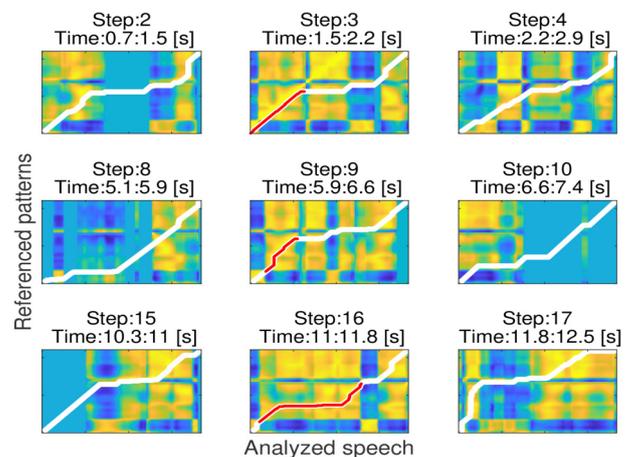


Fig 3. Operation of the discussed method presented on the selected fragments of analyzed speech. White stripes show optimal alignment paths between referenced pattern and analyzed fragment. Red stripes show the best matches selected after classification stage.

⁴ Only unique values were observed.

During performing the experiments using TTS query, some method variables have been recalculated, although during experiments with different sessions of the corpus, all variables haven't been changed.

IV. CONCLUSION

Results of the work presented in this paper are satisfactory, but the overall performance, as comparing to the original application of the method to the English language [6], is lower (especially for TTS-generated queries), which shall be further investigated. Possible improvement of the performance the author sees in employing formant frequencies analysis in the verification step of the method, as it is described in [7].

Additional study on TTS generation for the Polish language and its influence to the detecting properties of the method shall also be further investigated.

The method has many variables which are depended on the analyzed data. The optimization of the variables values has to be done according to applications.

REFERENCES

- [1] J. Sas, A. Żołnierczyk, "Pipelined language model construction for Polish speech recognition" in *International Journal of Applied Mathematics and Computer Science*, vol. 23, no. 3, 2013, pp. 649-668, DOI: 10.2478/amcs-2013-0049
- [2] D. Korżinek, Ł. Brocki, "Grammar Based Automatic Speech Recognition System for the Polish Language" in R. Jabłoński, M. Turkowski, R. Szewczyk (eds), *Recent Advances in Mechatronics*, Springer, Berlin, Heidelberg, 2007, ISBN 978-3-540-73956-2, pp. 87-91, DOI: 10.1007/978-3-540-73956-2_18
- [3] M. Ziółko, J. Gałka, B. Ziółko, T. Jadczyk, et al., "Automatic Speech Recognition System Dedicated for Polish" in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2011*, pp. 3315-3316.
- [4] B. Ziółko, T. Jadczyk, D. Skurzok, P. Zelasko, et al, *SARMATA 2.0 Automatic Polish Language Speech Recognition System*, Conference: *Interspeech 2015*, Dresden, Germany, 2015.
- [5] M. Pol, T. Walkowiak, M. Piasecki, "Towards CLARIN-PL LTC Digital Research Platform for: Depositing, Processing, Analyzing and Visualizing Language Data" in I. Kabashkin, I. Yatskiv, O. Prentkovskis (eds), *Reliability and Statistics in Transportation and Communication, RelStat 2017, Lecture Notes in Networks and Systems*, vol 36, Springer, Cham, 2018, pp. 485-494, DOI: 10.1007/978-3-319-74454-4_47.
- [6] Ł. Laszko, "Word detection in recorded speech using textual queries", *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, 2015, pp. 849-853, DOI: 10.15439/2015F341.
- [7] Ł. Laszko, "Using formant frequencies to word detection in recorded speech", *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, 2016, pp. 797-801, DOI: 10.15439/2016F518.
- [8] D. Korżinek, K. Marasek, Ł. Brocki, K. Wolk, "Polish Read Speech Corpus for Speech Tools and Services", *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26-28 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, number 136, Linköping University Electronic Press, Linköpings universitet, 2017, pp. 54-62.

Soccer Object Motion Recognition based on 3D Convolutional Neural Networks

Jiwon Lee, Do-Won Nam, and Wonyoung Yoo

SW-Content Research Laboratory,

Electronics and Telecommunications Research Institute,
218 Gajeong-ro, Yuseong-gu, Daejeon, Republic of Korea
Email: {ez1005, dwnam, zero2}@etri.re.kr

Yoonhyung Kim, Minki Jeong, and Changick Kim

Electrical Engineering,

Korea Advanced Institute of Science and Technology,
291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea
Email: {yhkim, rhm033, changick}@kaist.ac.kr

Abstract—Due to the development of video understanding and big data analysis research field using deep learning technique, intelligent machines have replaced the tasks that people performed in the past in various fields such as traffic, surveillance, and security area. In the sports field, especially in soccer games, it is also attempting quantitative analysis of players and games through deep learning or big data analysis technique. However, because of the nature of soccer analysis, it is still difficult to make sophisticated automatic analysis due to technical limitations. In this paper, we propose a deep learning based motion recognition technique which is the basis of high level automatic soccer analysis. For sophisticated motion recognition, we maximize recognition accuracy by sequentially processing the data in three steps: data acquisition, data augmentation, and 3D CNN based motion classifier learning. As can be seen from the experimental results, the proposed method guarantees real-time speed performance and satisfactory accuracy performance.

I. INTRODUCTION

IN the past, professional sports field was a human-oriented area. The training of the player has been done through the subjective guidance based on the know-how and experience of the manager and the coaching staff. Even in the case of a game judgement, it is judged through the intuition and observation of the referee, and the occasional misjudgement by the referee is accepted as part of sports. In addition, sports audiences were able to enjoy sports through unilateral delivery of sports contents. However, in recent years, many changes have been made in the field of professional sports as a result of quantitative analysis of sports through sports science and ICT technology. The manager and coaching staff can use data and video-based match analysis tools (eg, dartfish video analysis tool [1]) to check the objective player performance or conditions in detail, and to enable player training method or tactical changes. It also uses technology to help referee judges such as high-speed camera readings (eg, hawk-eye technology [2]) and produces interesting content using brilliant visualization tools (eg, freeD technology in NFL [3]) to give a sense of sports immersion. These sports analytic technologies are being developed to reflect the needs of people in many directions, thus the sports analysis market size was \$4.7 Billions in 2017 [4].

This trend has also affected the professional soccer market. Germany World Cup is to take advantage of big data analytics company, SAP's Match Insights technology to improve the

home team performance and analyze the strengths and weaknesses of the away teams to win the 2014 World Cup [5]. In addition to SAP, many international companies such as Chyronhego, OPTA, Deltatre, GPSports, and StatSport have technologies and services to perform quantitative analyzes on soccer matches and players.

In general, quantitative analysis of soccer game is consist of three steps: multi-object tracking, event analysis, and tactical analysis. Multi-object tracking can be automatically performed due to technological advances. However, in the cases of event analysis and tactical analysis, which require understanding of high level semantic from a given match, data is still extracted depending on the manual work of the expert group, and only the big data extracted by hand is secondarily processed and visualized. There are many reasons why these steps are not automated, but one of the biggest reasons is that the soccer event can be recognized only by the motion information of the player or referee. For example, it is necessary to be able to recognize a tackle motion of the player, a movement of the head referee's hand, and a flag motion of the assistant referee so that the tackle event, the foul event, and the offside event can be recognized. To solve this problem, this paper proposes a soccer object motion recognition technique.

This paper is composed as follows. In Sec. II, we describe the related researches. In Sec. III, we propose a soccer object motion recognition pipeline based on 3D convolutional neural networks (CNN). Sec. IV shows the experimental results of the proposed method. Finally, Sec. V discusses the concluding remarks.

II. RELATED WORKS

Motion recognition is a kind of computer vision field that recognizes human pose or action. The general process of motion recognition is as follows: 1) extracting feature points necessary for motion recognition in a given input source; 2) analyzing pattern of obtained feature points; 3) calculating similarity with predefined motion list; and 4) determining the final motion that has the highest similarity for the given input source. It is a kind of image classification technology in that the purpose of video-based motion recognition technology is to determine the final motion based on the similarity with predefined motion list.

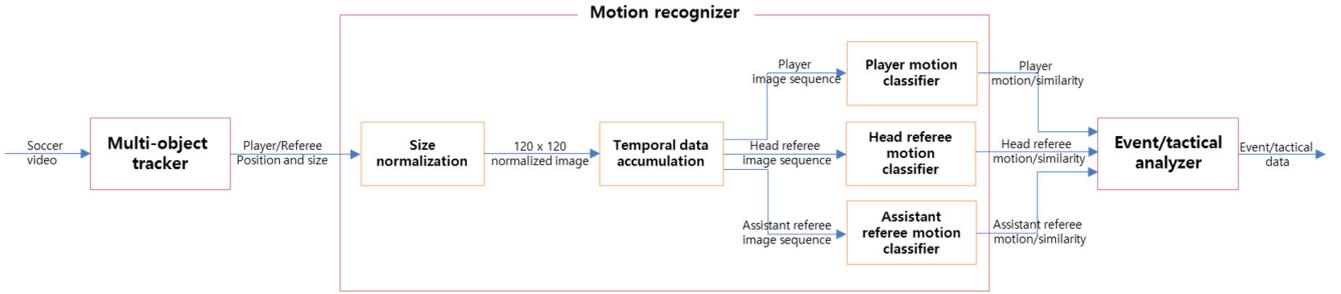


Fig. 1. System outline of the proposed method

Conventional motion recognition technologies are divided into several subdivisions according to several criteria. As a first criterion, it can be classified into two-dimensional(2D) motion recognition and three-dimensional(3D) motion recognition according to the dimension of the input source. 2D motion recognition technique performs motion recognition from 2D video sources taken from a general camera equipment [6], [7], [8]. 3D motion recognition technique performs motion recognition with stereoscopic video sources taken using special equipment such as MicroSoft's Kinect [9], [10]. As a second criterion, it can be classified as recognizing human action according to the human pose recognition and recognizing a gesture of a specific part of the human body. Motion recognition technique based on human pose tries to recognize motions such as human arm movements, arm extension, waist bending, and jumping motion based on video sources of human action [6], [7], [8], [9], [10]. Motion recognition technique for a specific gesture recognizes a partial movement of a specific part of the human body (hands, legs, *etc.*) [11], [12]. The third criterion is feature extraction method for motion recognition, and it is divided into hand-crafted feature extraction method and data-driven feature extraction method. A feature point is a clue that is used to distinguish different labels when performing motion classification. The accuracy of motion classification depends on the quality of the feature points. The hand-crafted feature extraction method is a method in which the user manually designs and extracts feature points according to a given classification purpose [6], [7], [8], [9], [10], [11], [12]. The hand-crafted feature extraction method are advantageous in that direct design of the user is easy and the patterns of motions to be classified are monotonous, but they have a disadvantage in that the performance is significantly lowered for motions with complex patterns. Recently, data-driven feature extraction method automatically learns feature points necessary for classification based on given information (video clip and label) [13]. Although this feature extraction method requires a large amount of computation and huge input data for learning, it performs much better than the hand-crafted feature extraction method in terms of accuracy and execution speed.

According to the above classification criteria, we can specify the category of motion recognition technique needed to solve

the problem defined in this paper. In this paper, we use 2D video sources taken from camera equipment installed in the stadium. The target area of the field player and referees in the game is tracked, and the goal is to recognize the motion based on the tracking data. In addition, it is possible to construct large-sized learning data, which is suitable for data-driven feature learning and extraction. According to this analysis, the motion recognition technology proposed in this paper can be specified as 1) 2D video source based, 2) data-driven feature extraction, and 3) technology to recognize human pose.

III. PROPOSED METHOD

In this Section, a method of performing motion recognition by inputting object regions tracked from a soccer game video will be described in detail. Figure 1 depicts the system outline of the proposed method. For the motion recognition specialized for the soccer object, we constructed the motion recognition system through three steps of data acquisition, data processing, and motion classifier learning [14]. A detailed description of each is given in the subsection.

A. Data Acquisition

The data acquisition is performed first to recognize the motion. To do this, we need to define motion classification criteria. We classify motions of each soccer object and generate learning data based on the following principles:

- The object is categorized into field player, head referee and assistant referee.
- All the motions that each object can take on the field must be included in the motion list.
- The body direction of the object with respect to the same motion secures data of at least four directions.

TABLE I
DEFINED MOTION LIST FOR EACH SOCCER OBJECT

Field player	Head referee	Assistant referee
Stand		Side
Walk	Walk	Walk
Run in	Run	Run
Kick	One arm pointing	Flag up
Tackle/Lie	Card	Flag chest
Throw in		Flag side

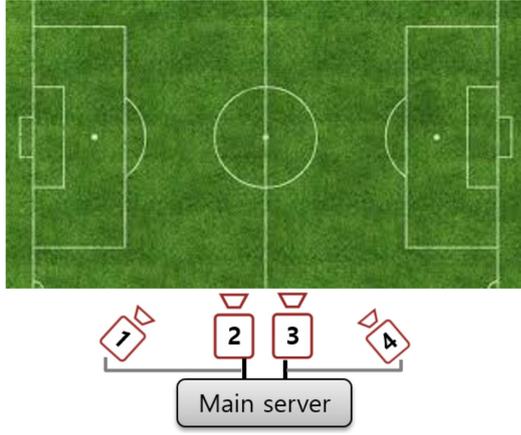


Fig. 2. Location of cameras in the stadium for data acquisition

- Motion with a duration of less than one second is excluded.

The motion classification list based on the above principles is shown in Table I.

After that, we acquired match videos through four cameras installed on a soccer stadium as shown in Fig. 2. In the figure, the first and fourth cameras shot the right half and the left half of the stadium respectively, and the second and third cameras shot the whole stadium. Here, the reason why the video was taken at various angles is to increase the recognition rate of the flag motion of the assistant referee. Then, the data necessary for motion classifier learning are acquired based on the object position extracted through the multi-object tracker [15].

B. Data Augmentation

After acquiring initial motion learning data, we extend the scale of learning data through data augmentation [16]. Motion data augmentation is closely related to the stability of the motion classifier. In general, the input to the motion classifier is a bounding box image including a soccer object which is an output of the multi-object tracker. Due to the nature of the tracking algorithm, the size and position of the bounding box fluctuate irregularly. Such trembling may cause performance degradation of motion recognition results. Therefore, a technique for effectively processing and extending motion learning data is needed for robust motion classifier that are robust to changes in bounding box size and position. We design a data processing algorithm suitable for this problem and incorporate it into motion classifier learning.

The designed data processing scheme is shown in Fig. 3, and its operation is as follows. First, a given image is normalized to an image of 140 pixels in width and in height. Then, a random cropping is performed at a size of 112 pixels in the horizontal and vertical directions. This process is introduced to imitate the phenomenon that the position of the bounding box shakes irregularly in the tracker output. Then, image up-scaling is performed in the following three ways for the image obtained through the random cropping. The first of the

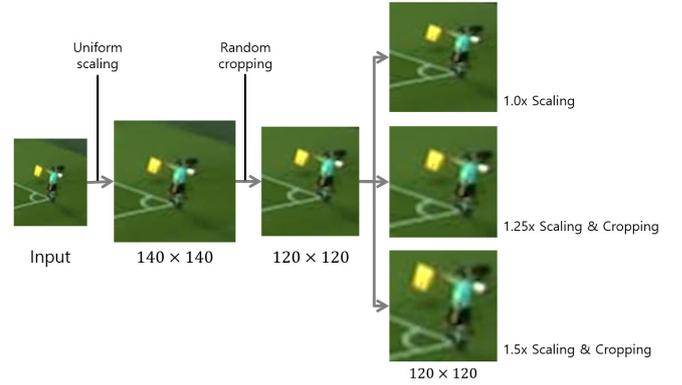


Fig. 3. Designed data augmentation scheme

three ways is to use the original image, and the second and third are to perform up-scaling of the pre-processed image by 1.25 times and 1.5 times, respectively, and then random cropping (with a size of 112 pixels in the horizontal and vertical directions). As shown in Fig. 3, as a result of this process, the relative sizes of the objects existing in the pre-processing image are divided into three, and three images are obtained from one original image. This process is called scale augmentation, which imitates that the size of the bounding box changes irregularly in the tracker output. We have learned to diversify the relative size of the object to be recognized through the scale augmentation so that the classifier can be robust to the perspective of the tracked object. In order to secure the robustness against trembling phenomenon of the tracking result, randomly cropped data was learned through random cropping. The learning data that has been processed through data augmentation process is finally used as input to the motion classifier after normalization process with a size of 120×120 pixels.

C. Motion Classifier Learning

Finally, we have learned a deep learning based motion classifier based on acquired and augmented learning data. A motion classifier performs learning according to a pre-determined number of labels at the training part and maps a given input image to one of the learned motion labels at the testing part. In this paper, we propose a 3D CNN-based motion classifier, which is a deep learning architecture that can understand the correlation between adjacent frames in order to take advantage of this feature, were used [17], [18].

Figure 4 shows a comparison of the 2D convolution and the 3D convolution. In the case of 2D convolution, a feature map

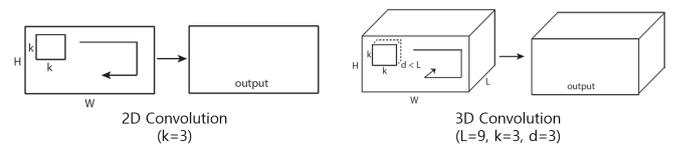


Fig. 4. Comparison of 2D convolution and 3D convolution

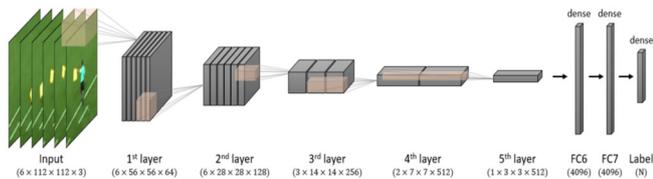


Fig. 5. Deep learning based motion classifier structure with 3D convolution (when $N = 6$)

is extracted using only spatial information for a single image, whereas a 3D convolution extracts not only spatial information but also temporal information for a plurality of continuous images to extract a feature map. Based on these features, the 3D CNN structure can be used to learn spatiotemporal information and contribute to performance enhancement and stabilization of the motion classifier. Figure 5 shows the network structure of a proposed motion classifier designed with a 3D CNN structure. The inputs to the network are F_b consecutive frame bundles (the frame bundle unit may vary depending on the applications), and a hierarchical feature map is extracted over a total of five layers for a given input video. In each feature map extraction step, 3D convolution is applied. A kernel having a differential depth (denoted by d in Fig. 5) is applied according to a layer of the feature map. The last three layers apply a fully connected network structure and apply a *softmax* function to finally output the similarity for N motions. Here, the *softmax* function is a generalization of the logistic function that normalizes a K -dimensional vector z having an arbitrary real values to a K -dimensional vector $\sigma(z)$ having real values in the range $[0, 1]$ with a sum of 1. The function is given by

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \text{ for } j = 1, \dots, K \quad (1)$$

IV. EXPERIMENTAL RESULTS

In order to verify the performance of the proposed motion classifier, we took $4K$ -sized videos at four different locations

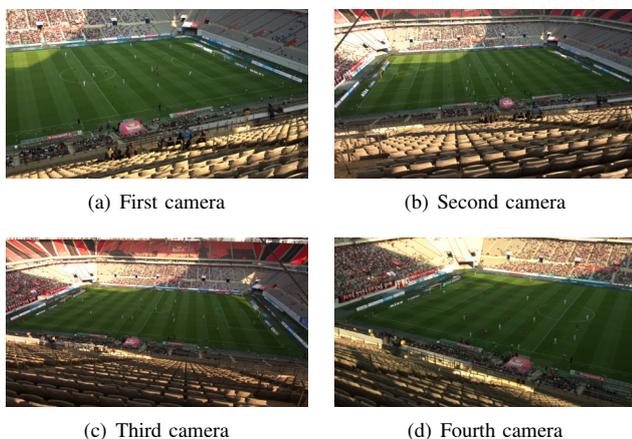


Fig. 6. A sample screenshot of four cameras in the test video clip



Fig. 7. An example of extracted data in each object

in the soccer stadium (see Fig. 2) for 9 K-league classic from 2016 to 2017. The captured video then proceeded to object tracking and divided the tracked data into field player, head referee, and assistant referee to generate learning data. A total of 170,000 pieces of initial learning data were generated, and about 600,000 pieces of final learning data were constructed after data augmentation process. Example screenshots of test videos and the generated learning data we have used are shown in Fig. 6 and Fig. 7, respectively.

In the case of soccer game, 3D CNN based motion classifiers are designed to enable parallel processing using tensorflow [19], since the number of objects appearing in one frame is large at the same time. In order to improve the accuracy of motion classifier, we need to consider not only spatial clue but also temporal clue. Here, we provide different temporal clues according to the characteristics of object to be recognized. In more detail, since the motion of the field player occurs with a shorter duration than the motion of the referee, the field player classifies the motion into 4 frames by one unit ($F_b = 4$), but in the case of the referee, 6 frames are grouped into one unit ($F_b = 6$) to perform motion classification.

To evaluate the performance of the proposed motion classifier, we used the i7-6770 core processor, DDR3 64GB RAM, and three different GPUs. The performance of the motion

TABLE II
PERFORMANCE VARIATIONS OF MOTION CLASSIFIER FOR EACH GPU

GPU types	Performance	
	ops	fps
GeForce GTX 1070	800	32
GeForce GTX TITAN X	930	37.2
NVIDIA TITAN X	1200	48

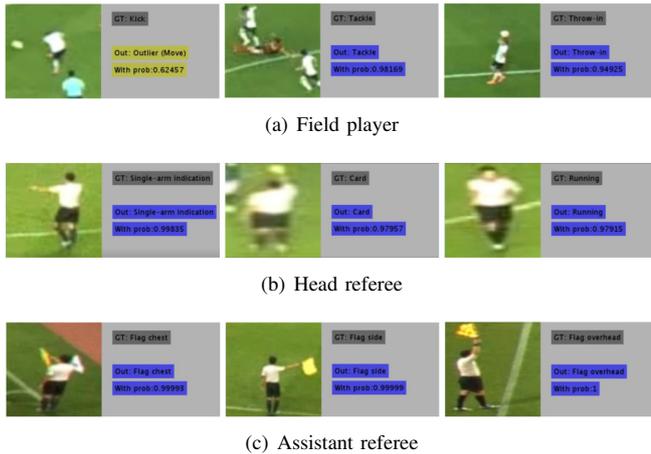


Fig. 8. Output examples of each motion classifier

classifier measured for each GPU is shown in Table II. Here, ops and fps refer to objects per second and frames per second, respectively. Since the number of field players and referees in one frame are 22 and 3, respectively, so the fps is calculated by dividing 25 from ops as shown in Table II. It means the proposed classifier has real-time motion recognition capability.

The finally obtained confusion matrix of each motion classifier is shown in Table III, IV, V, and the output example of the motion classifier is depicted in Fig. 8.

As can be seen from the experimental results, the average motion recognition accuracy of the field player, the head referee, and the assistant referee was 0.449, 0.851, and 0.872, respectively. It can be confirmed that the accuracy of the motion recognition of the field player is relatively low compared to the referee. In the case of the referees, it is easy to distinguish the motion because the number of motion to be recognized is small and the motion itself is stereotyped.

TABLE III
CONFUSION MATRIX OF MOTION CLASSIFIER FOR FIELD PLAYER

Out \ GT	Stand	Walk	Run	Kick	Tackle /Lie	Throw in
Stand	1,779	0	0	1	0	0
Walk	58	325	1,158	261	6	0
Run	0	0	0	4	0	0
Kick	569	694	702	1,440	1,147	648
Tackle/Lie	70	1	0	207	768	357
Throw in	9	254	37	308	35	989
Accuracy	0.716	0.255	0	0.648	0.393	0.496

TABLE IV
CONFUSION MATRIX OF MOTION CLASSIFIER FOR HEAD REFEREE

Out \ GT	Walk	Run	One arm pointing	Card
Walk	11,244	202	376	17
Run	1,210	10,579	446	132
One arm pointing	395	499	8,254	143
Card	388	295	402	546
Accuracy	0.850	0.914	0.871	0.652

TABLE V
CONFUSION MATRIX OF MOTION CLASSIFIER FOR ASSISTANT REFEREE

Out \ GT	Sidle	Walk	Run	Flag up	Flag chest	Flag side
Sidle	12,610	131	126	491	962	1,279
Walk	116	12,564	107	374	723	1,369
Run	51	159	11,887	216	430	342
Flag up	0	0	4	9,033	869	621
Flag chest	2	0	3	1,153	16,030	442
Flag side	46	18	5	1,616	1,317	11,724
Accuracy	0.983	0.976	0.980	0.701	0.788	0.743

However, in the case of the field player, the number of motions to be recognized is relatively large and the duration of occurred motion is also shorter than that of the referees (Recognition using only 66% frames compared to referees). In addition, the field player has a high degree of similarity between different motions, which is considered to have affected the accuracy.

V. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we introduced data acquisition, data processing, and motion classifier learning method to recognize motion of soccer objects from soccer video. In particular, to design motion classifiers with high accuracy, we use $3DCNN$, which is a structure that extracts spatio-temporal features well, and developed motion classifier considering real-time by using parallel processing technique. As can be seen from the experimental results, it can be seen that the proposed method satisfies the real-time speed performance and the high motion recognition accuracy of the referee. However, the accuracy of the recognition of field player motion is rather low, and further research is needed.

In the future, we will design a sophisticated motion classifier with high accuracy even for objects with ambiguous motion classification such as field player, and will try to incorporate the developed motion classifier into other sports fields such as basketball and figure skating.

ACKNOWLEDGMENT

This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Reasearch & Development Program 2016 (R2016030044, Development of Context-Based Sport Video Analysis, Summarization, and Retrieval Technologies)

REFERENCES

- [1] DartFish sports analysis tool [Online] Available : <http://www.dartfish.com>
- [2] Hawk-eye innovations [Online] Available: <https://www.hawkeyeinno vations.com>
- [3] FreeD on NFL [Online] Available : <https://newsroom.intel.com/news/intel-nfl-kickoff-freed-technology-11-stadiums-create-immersive-highlights-2017-season/>
- [4] "Sports analytics: market shares, strategies, and forecasts, worldwide, 2015 to 2021," Wintergreen Research, 472 pages, May 2015
- [5] A. Ghosh, "How 'Match Insight' is changing soccer," 6th Aug. 2014. [Online] Available: <https://blogs.sap.com/2014/08/06/how-software- is-making-football-even-more-beautiful/>

- [6] C. P. Huang, C. H. Hsieh, K. T. Lai, and W. Y. Huang, "Human action recognition using histogram of oriented gradient of motion history image," in *International Conference on Instrumentation, Measurement, Computer, Communication and Control*, pp. 353-356, Oct. 2011.
- [7] L. Hu, W. Liu, B. Li, and W. Xing, "Robust motion detection using histogram of oriented gradients for illumination variations," in *Proc. ICIMA 2010*, pp. 443-447, May. 2010.
- [8] P. Banerjee and S. Sengupta, "Human motion detection and tracking for video surveillance," in *National Conference for Communication*, 2008.
- [9] O. Patsadu, C. Nukoolkit, and B. Watanapa, "Human gesture recognition using Kinect camera," in *Proc. JCSSE 2012*, pp. 28-32, May. 2012.
- [10] E. E. Stone and M. Skubic, "Fall detection in homes of older adults using the Microsoft Kinect," *IEEE Jour. Biomedical and Health Informatics*, vol. 19, no. 1, pp. 290-301, Mar. 2014.
- [11] N. C. Kiliboz and U. Gudukbay, "A hand gesture recognition for human computer interaction," *Jour. Visual Communication and Image Representation*, vol. 28, pp. 97-104, Apr. 2015.
- [12] M. B. Brahem, B. J. Menelas, and M. D. Otis, "Use of 3DOF accelerometer for foot tracking and gesture recognition in mobile HCI," *Peocedia Computer Science*, vol. 19, pp. 453-460, 2013.
- [13] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," in *Nature*, vol. 521, pp. 436-444, May. 2015.
- [14] J. Lee, Y. Kim, M. Jeong, C. Kim, D. Nam, J. Lee, S. Moon, and W. Yoo, "3D convolutional neural networks for soccer object motion recognition," in *Proc. ICACT 2018*, pp. 354-358, Feb. 2018.
- [15] W. Kim, S. Moon, J. Lee, D. Nam, and C. Jung, "Multiple Player Tracking in Soccer Videos : An Adaptive Multiscale Sampling Approach," *Multimedia Systems*, pp. 1-13, Feb. 2018.
- [16] A. Krizhevsky, I. Sutskever, and G. E. hinton, "ImageNet classification with deep convolutional neural network," in *Proc. NIPS 2012*, pp. 1-9, Dec. 2012.
- [17] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, Mar. 2012.
- [18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. ICCV 2015*, pp. 4489-4497, Dec. 2015.
- [19] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467v2*, 2016.

Analysis of inter-channel dependencies in audio lossless block coding

Cezary Wernik
West Pomeranian University of
Technology in Szczecin,
al. Piastów 17, 70-310 Szczecin,
Poland
Email: cwernik@wi.zut.edu.pl

Grzegorz Ulacha
West Pomeranian University of
Technology in Szczecin,
al. Piastów 17, 70-310 Szczecin,
Poland
Email: gulacha@wi.zut.edu.pl

Abstract— In this paper the basics of data predictive modeling (using the method of minimization mean square error) for lossless audio compression are presented. The described research focuses on inter-channel analysis and setting range of prediction in dependencies of frame size. In addition, the concept of data flow using inter-channel dependencies and an authorial, effective and flexible method of saving prediction coefficients are presented.

I. INTRODUCTION TO LOSSLESS AUDIO DATA COMPRESSION

MINIMIZATION of storage and transmission data cost are one of most important issues of teleinformatics. Tool for simplification of reduction of this cost is data compression. A lot of such compression algorithms exist and the most effective ones are adapted to the specific data type. Compression methods can be divided into lossy and lossless, and this research focuses on the latter, limited to the coding of audio data.

Important purposes of lossless audio compression include recording storage, saving of records with high-quality sound on commercial media (e.g. DVDs, Blu-Ray), and selling songs in online music stores for more demanding customers who are not satisfied with the quality of mp3 format [5]. Moreover, lossless mode is often required at the stage of music processing in a studio, advertising materials and in the production of radio and television programs, films (post-production [1]) etc. In such case, no lossy coding is used, which at each iteration of sound editing may cumulate additional distortions.

In the beginning of the 21st century, many effective proposals for the MPEG-4 Lossless Audio Coding standard were developed [4], but we cannot ignore the fact that a branch of amateur solutions develops in an independent way, which use algorithms that are not fully presented in scientific publications. For example OptimFrog [14] and Monkey's Audio [15] belong to the top of most efficient programs for lossless audio compression.

In modern compression methods usually two steps are used: data decomposition, and then compression by one of the efficient entropy methods. The most effective ones are arithmetic coding, Huffman coding [10] and its variations

such as Golomb [3] and Rice [8] code. In the case of audio coding, a Gilbert-Moore block code, which is a variation of the arithmetic code [7] is also used.

Publications in the field of lossless compression of audio are mainly focused on the stage of data decomposition. There are two basic types of modeling. The first is the use of linear prediction [2], [9] or non-linear (e.g. using neural networks [6]). The second type is the use of such transformations as DCT (MPEG-4 SLS [12]) or wavelet, but the current research shows that this type is slightly less efficient in the case of lossless coding. Therefore, predictive methods are used in most cases.

In Sections II and III the basics of audio modeling, including calculation (based on the mean square error minimization method) and a typical method of saving prediction coefficients are presented. In Section IV the authorial solution of flexible and effective method saving prediction coefficients are presented, afterwards In Section V presented the analysis of inter-channel dependencies. In Section VI was described the effective way of coding prediction errors (Golomb adaptive code), while in Section VII presented the summary and comparison of the efficiency of the proposed method against other known solutions.

II. BASICS OF AUDIO MODELING

In lossless audio codec's, typical linear predictor of the order r is used for modeling, which is the predicted value of the currently coded sample $x(n)$ based on weighted average of the r previous signal samples. The simplest predictive models are those with fixed coefficients, including, the DPCM constant model using the previous sample $\hat{x}(n) = x(n-1)$. The use of a linear predictor allows to encode only prediction errors, i.e. differences $e(n)$ between the actual and predicted value (rounded to the nearest integer), which are most often small values oscillating near zero:

$$e(n) = x(n) - \hat{x}(n) \quad (1)$$

In this way we obtain a difference signal in which the distribution of errors $e(n)$ has a character similar to Laplace distribution, which allows for efficient coding using one of

static or adaptive entropy methods, such as Golomb-Rice code or arithmetic code (see Fig. 1 in chapter V).

Usually a method called forward adaptation is used, as the encoder must have access to the whole frame before encoding, and this means that the calculated coefficients should also be sent to the decoder. For this reason, when developing the method with forward adaptation, one should calculate the effective way of choosing the frame size, the order of prediction, as well as the accuracy of saving the prediction coefficients. This is an asymmetrical method temporarily, since the decoder works relatively quickly, downloading only the head information associated with the given frame and decoding based on the formula (1). The disproportion of time in the method of forward adaptation (characterized by a longer coding time in relation to decoding) plays a significant role, since the coding operation is most often carried out once, and decoding many times.

A typical forward adaptation solution is used in the MPEG-4 ALS, where the frame is approximately 43 ms long (depending on the sampling frequency, the frame length counted in the samples is different and the maximum at $f_p = 192$ kHz is $N = 8192$).

We can also introduce the term *long term* of the frame (which is a group of frames) called *super-frame* here, in which the number N can be hundreds of thousands of samples. Although MPEG-4 ALS uses frames with a maximum size of $N = 2^{13}$, there is no obstacle in increasing length of frames in newer solutions. Their length may be limited by the principle of free access to data in real time, which results from the needs of e.g. studio sound processing. For this purpose, MPEG-4 ALS proposes independent access to the frame set every 500 ms (no need to decode previous data to correctly decode any frames), which introduces a limit of $N_{\max} = 24\ 000$ samples in one super-frame when sampling frequency is 48 kHz. Above this value one should give up the possibility of free access. However, this is not a problem for archiving and transmission applications, e.g. when someone wants to purchase the whole artist's album from the online store.

In this paper, we consider the application of super-frames with a length of 20 seconds. The analysis was made on 16 dozen-second fragments of recordings (stereo, 16-bit samples, 44 100 samples per second) of various genres of music, men's and women's speech recordings available in base [16].

III. CALCULATION OF PREDICTION COEFFICIENTS USING MMSE METHOD

It has been widely accepted that for audio signals, the calculation of prediction coefficients using the Mean Square Error Minimization (MMSE) method gives very good results.

To calculate the prediction coefficients in practice, the Autocorrelation Levinson-Durbin method is most often used [13], which by simplification does not require the calculation

of the inverse matrix, but it is able to calculate the model coefficients in an iterative manner for subsequent orders of prediction. In this way, we reduce computational complexity from $O(n^3)$ to $O(n^2)$. An additional advantage is that the reflection coefficients $\mathbf{k} = [k_1, k_2, \dots, k_r]^T$ often referred to as PARCOR (partial correlation coefficients) belong to the compartment $(-1; 1)$ and they can be effectively coded (they are subjected to a quantization process, e.g. using 7 bits). Using this, the size of the header containing the set of coefficients of a given model is not large, even if quite high orders of prediction are used [5].

In contrast to the PARCOR format, which assumes the stationarity of the audio signal the new approach proposed allows for coding efficiently prediction coefficients obtained in any way (e.g. thanks to the use of different types of suboptimal algorithms that minimize the entropy value or a bitwise average in each coded frame) [10].

Rejecting the assumption about the stationarity of audio signal should be used the autocovariance method, wherein the vector of prediction coefficients $\mathbf{w} = [w_1, w_2, \dots, w_r]^T$ is calculated from the matrix equation [10]:

$$\mathbf{w} = \mathbf{R}^{-1} \cdot \mathbf{p} \quad (2)$$

where \mathbf{R} is a square matrix with dimensions $r \times r$ elements $R_{(j,i)}$ such that:

$$R_{(j,i)} = \sum_{n=0}^{N-1} x(n-i) \cdot x(n-j), \quad (3)$$

while \mathbf{p} is a vector with size $r \times 1$ elements $p_{(j)}$ such that:

$$p_{(j)} = \sum_{n=0}^{N-1} x(n) \cdot x(n-j), \quad (4)$$

where N is the number of samples in frame. It is assumed that for the first r samples in the first frame of super-frame a simplified prediction model of the lower order is used.

IV. ENCODING PREDICTION COEFFICIENTS METHOD

The method of saving header data proposed in this work assumes that from the set of prediction coefficients \mathbf{w} we choose the one with the highest absolute value. We mark his position in the vector as i_{\max} , and the value of this coefficient as w_{\max} . Value of w_{\max} is initially projected onto a 32-bit float type, and the index i_{\max} is saved as an 8-bit integer (at $r \leq 256$).

All other coefficients are coded on $b + 1$ bits after normalizing their value in relation to w_{\max} and appropriate scaling, in a manner consistent with the formula:

$$\bar{w}_i = \left\lfloor \left(\frac{w_i}{w_{\max}} + 1 \right) \cdot \left(2^b - \frac{1}{2} \right) + \frac{1}{2} \right\rfloor, \quad (5)$$

Reconstruction of the original prediction coefficients from the encoded header follows the use of the formula:

$$w_i = \left(\frac{w_i}{2^b - \frac{1}{2}} - 1 \right) \cdot w_{\max}. \quad (6)$$

The length of the header in bits for each frame is calculated by formula $\text{float}_{\text{size}} + \lceil \log_r r \rceil + (r-1) \cdot (b+1)$, where $\text{float}_{\text{size}}$ is the size of a variable compliant with the standard float32, intended for saving the w_{\max} coefficient, which accuracy was experimentally reduced from 32 to 21 bits.

Theoretically, the higher the order of prediction we use, the more effective the prediction model we will get. Unfortunately higher r is the reason for the increase in the size of the frame header. At the same time it can be noticed that the increase in the order of predictions also increases the required accuracy of the vector \mathbf{w} coefficients. Both of these parameters indicate that it may be more profitable to use smaller order thus reducing the size of the frame. If frames are shorter, the more accurate matching of predictive models to the variability in time of the audio characteristics is. On the other hand, too short frames cause significant increase of the header, which in turn leads to the need to further reduce the order of prediction and the value of b . Based on conducted experiments, specific b values are set for ranges of prediction orders r , at given frame lengths 2^q as presented in Table I.

TABLE I.

OPTIMAL b -VALUES RELATIVE TO THE ORDER OF PREDICTION r WITH 2^q FRAME LENGTH

Q	b	
9	11, when $r < 24$	10, when $r \geq 24$
10	11	
11	12, when $r < 12$	11, when $r \geq 12$
12	12, when $r < 192$	11, when $r \geq 192$
13 and 14	12	

V. EXPLOITING INTER-CHANNEL DEPENDENCIES

Existing dependencies between channels allows to use in r -predictive models samples from a both channels, left $x_L(n-i)$ and right $x_R(n-j)$.

$$\begin{aligned} \hat{x}_L(n) &= \sum_{i=1}^{r_L} a_i^{(L)} \cdot x_L(n-i) + \sum_{j=1}^{r_R} b_j^{(L)} \cdot x_R(n-j), \\ \hat{x}_R(n) &= \sum_{j=1}^{r_L} b_j^{(R)} \cdot x_R(n-j) + \sum_{i=0}^{r_R-1} a_i^{(R)} \cdot x_L(n-i) \end{aligned} \quad (7)$$

There are vectors of prediction coefficients for the left channel $\mathbf{w}_L = [a_1^{(L)}, a_2^{(L)}, \dots, a_{r_L}^{(L)}, b_1^{(L)}, b_2^{(L)}, \dots, b_{r_R}^{(L)}]^T$ and $\mathbf{w}_R = [b_1^{(R)}, b_2^{(R)}, \dots, b_{r_L}^{(R)}, a_0^{(R)}, a_1^{(R)}, \dots, a_{r_R-1}^{(R)}]^T$ the right channel. Fig. 1. shows the data flow and existing dependencies between them. The formulas are two because by coding (decoding) the value of the right channel sample

$x_R(n)$, there is already access to the current sample of the left channel $x_L(n)$.

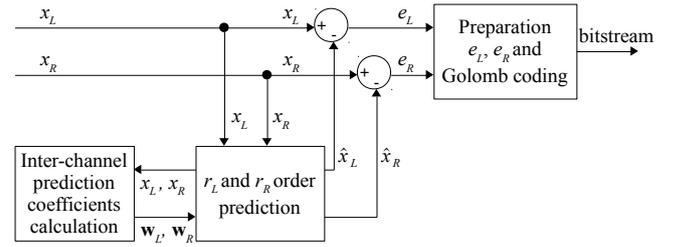


Fig. 1 Schema of proposition our algorithm

The result of the bit average can be influenced by the selection of which channels are coded in the first order. It should be clearly noted that the r_L ordering in both cases concerns the set of samples of the currently coded channel, while r_R is number samples of the opposite channel, in addition $r = r_L + r_R$.

There is a problem with the selection of the universal r_R/r_L ratio. The most common proportions in the literature are 1:1 and 1:2, but after completing the bit-minimizing experiments it turned out that depending on the size of frame (2^q), the best order of prediction r is changing (within the test database). Also the increase order of the prediction r leads to a decreasing value of the ratio r_R/r_L . The results of experiments are presented in Table II.

TABLE II.

EXAMINATION OF THE OPTIMAL RATIO r_R/r_L IN RELATION TO THE LENGTH OF THE 2^q FRAME

q_r	R	r_L	r_R	$\frac{r_R}{r_L}$	L_{avg}
9	12	6	6	1,000	9,369
10	15	8	7	0,875	9,249
11	19	11	8	0,727	9,202
12	34	22	12	0,545	9,246
13	58	42	16	0,381	9,335
14	57	44	13	0,295	9,438

On Fig. 2. was show the average level of Pearson's correlation coefficient (for the whole test base), using the best settings $\{q, r_L, r_R\} = \{11, 11, 8\}$, between the coded sample $x_L(n)$ and 11 adjacent samples $x_L(n-i)$ left channel and 8 samples (graph bars (lag) with indexes 12 to 19) of the right channel $x_R(n-j)$. The shape of the chart for individual files does not differ significantly from the one shown in Fig. 2. This can not be said about the charts of the average level of absolute value of prediction coefficients, which differ significantly for individual files. The ATrain and TomsDiner files were selected from the test database, for which the average level of absolute values of prediction coefficients are presented in Fig. 3 and Fig. 4, respectively. For both of these files, a bit average analysis was made depending on the proportion of r_L to r_R with a constant value

of $r = 19$ (see Fig. 5). The best bit averages was obtained with $\{r_L, r_R\} = \{18, 1\}$ for file ATrain and with $\{r_L, r_R\} = \{7, 12\}$ for file TomsDiner. However, these optimal settings can not be deduced only from analysis Fig. 2-4, and the procedure for scanning all settings $r = 19$ (resulting in the data from Fig. 5) leads to a significant complexity of the implementation of the encoder. However, these optimal settings can not be deduced solely from based on the analysis of Fig. 2-4. For this reason, use the scanning procedure all settings ($r=19$) lead to a significant complexity of the implementing coder. This shows that the proportions r_L to r_R differ significantly from the common for the whole base of the best compromise pair $\{r_L, r_R\} = \{11, 8\}$.

VI. GOLOMB CODE APPLICATION FOR PREDICTION ERROR CODING

Golomb code [3] is a specific version of the Huffman code [10] which is a prefix code for a source with an infinite symbol alphabet. It is used to represent integers consistent with the geometric distribution. It works well for encoding audio after their predictive modeling.

The main advantage of the Golomb code is that it does not require the use of code tables and is relatively simple in hardware implementation. Golomb's code with forward adaptation was used in this work, calculating the individual parameter m of this code for each frame with a length of 512 samples. Its saved on 13 bits is sufficient to describe the local probability distribution.

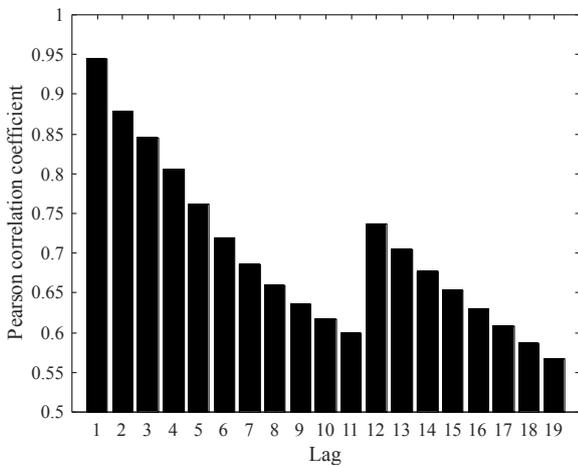


Fig. 2 Inter-channels mean of absolute value Pearson correlation coefficients

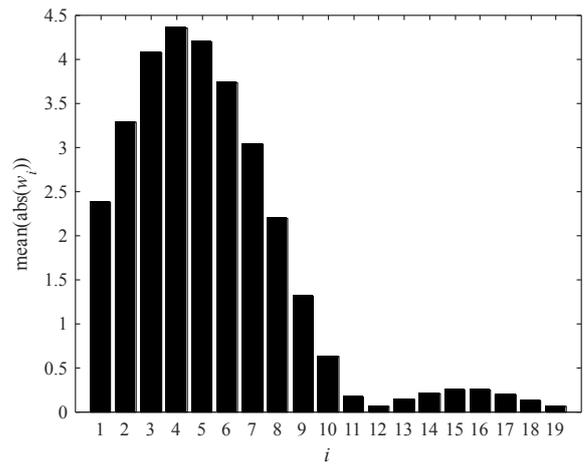


Fig. 3 Mean value of absolute prediction coefficients for ATrain file

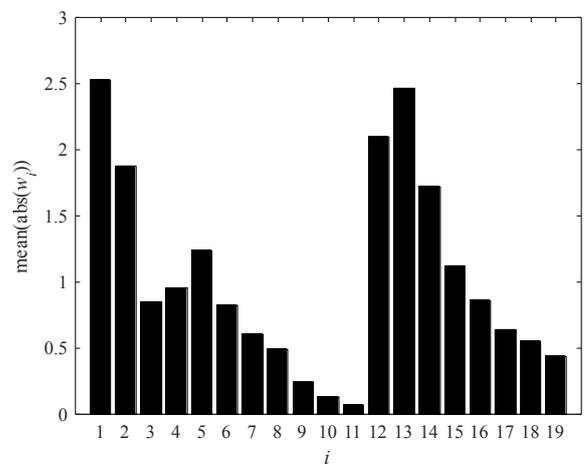


Fig. 4 Mean value of absolute prediction coefficients for TomsDiner file

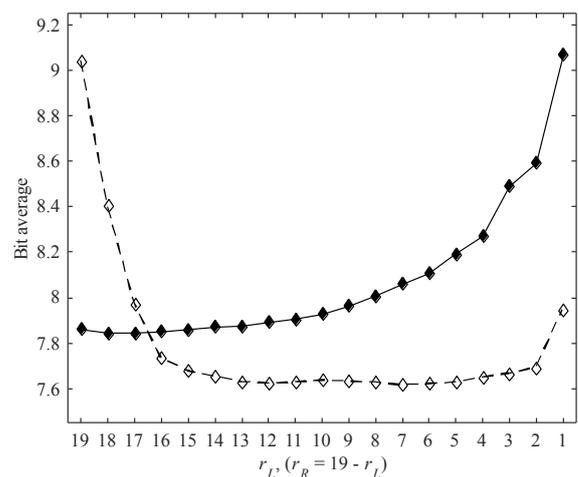


Fig. 5 Bit average for ATrain (continuous line) and TomsDiner (long dashed line) files in dependencies of r_L to r_R proportions

VII. SUMMARY

Using the proposed method (MMSE) for the universal parameter set $\{q, r_L, r_R\} = \{11, 11, 8\}$ the bit average for the test database (used in [11]) was better by 14.95% compared to the universal archiving tool RAR 5.0, as well as 11.81% better than the dedicated SHORTEN 3.6.1 solution. In compare to default mode of MP4-ALS-RM23 our algorithm are better about 2.11%, but still, the results are worse than the best published solutions, such as MP4-ALS-RM23 turned in the best (but slow) mode by using special switches, which is presented in Table III.

TABLE I.
COMPARISON OF CODES FOR BASE 16 AUDIO FILES [11]

Codec	RAR 5.0	Shorten 3.6.1	MP4-ALS-RM23 (default mode)	WavPack 5.1 [17]	Our proposition	MP4-ALS-RM23 (the best mode)
bit average	10.820	10.434	9.352	9.208	9.202	8.718

The main purpose of this work was to show directions of optimization of predictive models, such as choosing the order of coded channels or individual r_R/r_L ratio (for a given file or even each frame).

In the MP4-ALS-RM23 standard, not everything has been fully developed. The proposal of our algorithm is similar in construction to MP4-ALS in basic ALS version without RLSLMS and BGMC mode, but building the next steps of the algorithm we fill the gaps that were omitted in MP4, which will increase overall efficiency in the future. Our algorithm loses with MP4 (the best mode in Table III) because we tested MP4 in the best mode using switches: -z3, -p, -b, providing respectively: RLSLMS mode (in best configuration), long-term prediction, using BGMC codes for prediction residual (instead Rice code). In the current version of our algorithm, using the Golomb code we have a faster implementation than we would use arithmetic coding used in MP4 with BGMC mode. MP4-ALS-RM23 have many switches, which were chosen best for effective compression. Our proposal has a static configuration that is flexible ad-hoc.

Further work using the approach proposed in this paper, with the use of ideas employed, among others, in [11] (introducing cascading combining of successive blocks to minimize prediction errors) will allow to significantly

improve the efficiency of forward coder. It is also expected to introduce a better than MMSE technique for the selection of prediction coefficients as well as a more accurate individual selection of the parameters $\{q, r_L, r_R\}$ described here for each super-frame, as in case of MPEG4-ALS.

REFERENCES

- [1] S. Andriani, G. Calvagno, T. Erseghe, G. A. Mian, M. Durigon, R. Rinaldo, M. Kneé, P. Walland, M. Koppetz, Comparison of lossy to lossless compression techniques for digital cinema, *Proceedings of International Conference on Image Processing ICIP'04*, 24-27 Oct. 2004, vol. 1, pp. 513-516.
- [2] C. D. Giurcaneau, I. Tabus, J. Astola, Adaptive context based sequential prediction for lossless audio compression, *Proceedings of IX European Signal Processing Conference EUSIPCO 1998*, Rhodes, Greece, Sept. 1998, vol. 4, pp. 2349-2352.
- [3] S. W. Golomb, Run-length encoding, *IEEE Transactions on Information Theory*, July 1966, vol. 12, pp. 399-401.
- [4] H. Huang, P. Fränti, D. Huang, S. Rahardja, Cascaded RLS-LMS prediction in MPEG-4 lossless audio coding, *IEEE Trans. on Audio, Speech and Language Processing*, March 2008, vol. 16, no. 3, pp. 554-562.
- [5] T. Liebchen, Y. A. Reznik, Improved Forward-Adaptive Prediction for MPEG-4 Audio Lossless Coding, in *118th AES Convention*, 28-31 May 2005, Barcelona, Spain, pp. 1-10.
- [6] E. Ravelli, P. Gournay, R. Lefebvre, A Two-Stage MLP+NLMS Lossless coder for stereo audio, *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, 14-19 May 2006, vol. 5, pp. V_177-180.
- [7] Y. A. Reznik, Coding of prediction residual in MPEG-4 standard for lossless audio coding (MPEG-4 ALS), *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, Montreal, Quebec, Canada, 17-21 May 2004, vol. 3, pp. III_1024-1027.
- [8] R. F. Rice, *Some practical universal noiseless coding techniques*, Jet Propulsion Laboratory, JPL Publication 79-22, Pasadena, CA, March 1979.
- [9] T. Robinson, SHORTEN: Simple lossless and near-lossless waveform compression, Cambridge Univ. Eng. Dept., Cambridge, UK, Tech. Rep. 156, 1994, pp. 1-17.
- [10] K. Sayood, *Introduction to Data Compression*, 2nd edition, Morgan Kaufmann Publ., San Francisco, 2002.
- [11] G. Ulacha, R. Stasiński, Entropy Coder for Audio Signals, *International journal of electronics and telecommunications*, Vol. 61, No. 2, Poland, pp. 219-224, 2015
- [12] R. Yu, S. Rahardja, C. C. Ko, H. Huang, Improving coding efficiency for MPEG-4 Audio Scalable Lossless coding, *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, PA, USA, 18-23 March 2005, vol. 3, pp. III_169-172.
- [13] Yuli You, *Audio coding. Theory and applications*, 1st edition, Springer, New York 2010.
- [14] <http://www.losslessaudio.org/>
- [15] <http://www.monkeysaudio.com/>
- [16] http://www.rarewares.org/test_samples/
- [17] <http://www.wavpack.com/>

International Conference on Innovative Network Systems and Applications

MODERN network systems encompass a wide range of solutions and technologies, including wireless and wired networks, network systems, services and applications. This results in numerous active research areas oriented towards various technical, scientific and social aspects of network systems and applications. The primary objective of Innovative Network Systems and Applications (iNetSApp) conference is to group network-related events and promote synergy between different fields of network-related research. To stimulate the cooperation between commercial research community and academia, the conference is co-organised by Research and Development Centre Orange Labs Poland and leading universities from Poland, Slovak Republic and United Arab Emirates.

The conference continues the experience of Frontiers in Network Applications and Network Systems (FINANS), International Conference on Wireless Sensor Networks (WSN), and International Symposium on Web Services (WSS). As in the previous years, not only research papers, but also papers

summarising the development of innovative network systems and applications are welcome.

- CAP-NGNCS'18—1st International Workshop on Communications Architectures and Protocols for the New Generation of Networks and Computing Systems
- INSERT'18 - 2nd International Conference on Security, Privacy, and Trust
- IoT-ECAW'18—2nd Workshop on Internet of Things—Enablers, Challenges and Applications
- WSN'18 - 7th International Conference on Wireless Sensor Networks

AREA SUPERVISORY COMMITTEE

- Awad, Ali Ismail, INSERT'18
- Furtak, Janusz, IoT-ECAW'18
- Hamrioui, Sofiane, CAP-NGNCS'18
- Ševčík, Peter, WSN'18

2nd Workshop on Internet of Things—Enablers, Challenges and Applications

THE Internet of Things is a technology which is rapidly emerging the world. IoT applications include: smart city initiatives, wearable devices aimed to real-time health monitoring, smart homes and buildings, smart vehicles, environment monitoring, intelligent border protection, logistics support. The Internet of Things is a paradigm that assumes a pervasive presence in the environment of many smart things, including sensors, actuators, embedded systems and other similar devices. Widespread connectivity, getting cheaper smart devices and a great demand for data, testify to that the IoT will continue to grow by leaps and bounds. The business models of various industries are being redesigned on basis of the IoT paradigm. But the successful deployment of the IoT is conditioned by the progress in solving many problems. These issues are as the following:

- The integration of heterogeneous sensors and systems with different technologies taking account environmental constraints, and data confidentiality levels;
- Big challenges on information management for the applications of IoT in different fields (trustworthiness, provenance, privacy);
- Security challenges related to co-existence and interconnection of many IoT networks;
- Challenges related to reliability and dependability, especially when the IoT becomes the mission critical component;
- Zero-configuration or other convenient approaches to simplify the deployment and configuration of IoT and self-healing of IoT networks;
- Knowledge discovery, especially semantic and syntactical discovering of the information from data provided by IoT;

The IoT conference is seeking original, high quality research papers related to such topics. The conference will also solicit papers about current implementation efforts, research results, as well as position statements from industry and academia regarding applications of IoT. The focus areas will be, but not limited to, the challenges on networking and information management, security and ensuring privacy, logistics, situation awareness, and medical care.

TOPICS

The IoT conference is seeking original, high quality research papers related to following topics:

- Future communication technologies (Future Internet; Wireless Sensor Networks; Web-services, 5G, 4G, LTE, LTE-Advanced; WLAN, WPAN; Small cell Networks...) for IoT,

- Intelligent Internet Communication,
- IoT Standards,
- Networking Technologies for IoT,
- Protocols and Algorithms for IoT,
- Self-Organization and Self-Healing of IoT Networks,
- Trust, Identity Management and Object Recognition,
- Object Naming, Security and Privacy in the IoT Environment,
- Security Issues of IoT,
- Integration of Heterogeneous Networks, Sensors and Systems,
- Context Modeling, Reasoning and Context-aware Computing,
- Fault-Tolerant Networking for Content Dissemination,
- Architecture Design, Interoperability and Technologies,
- Data or Power Management for IoT,
- Fog—Cloud Interactions and Enabling Protocols,
- Reliability and Dependability of mission critical IoT,
- Unmanned-Aerial-Vehicles (UAV) Platforms, Swarms and Networking,
- Data Analytics for IoT,
- Artificial Intelligence and IoT,
- Applications of IoT (Healthcare, Military, Logistics, Supply Chains, Agriculture, ...),
- E-commerce and IoT.

The conference will also solicit papers about current implementation efforts, research results, as well as position statements from industry and academia regarding applications of IoT. Focus areas will be, but not limited to above mentioned topics.

EVENT CHAIRS

- **Cao, Ning**, College of Information Engineering, Qingdao Binhai University
- **Furtak, Janusz**, Military University of Technology, Poland
- **Zieliński, Zbigniew**, Military University of Technology, Poland

PROGRAM COMMITTEE

- **Amanowicz, Marek**, Military University of Technology
- **Antkiewicz, Ryszard**, Military University of Technology, Poland
- **Chudzikiewicz, Jan**, Military University of Technology in Warsaw, Poland
- **Cui, Huanqing**, Shandong University of Science and Technology, China

- **Ding, Jianrui**, Harbin Institute of Technology, China
- **Fouchal, Hacene**, University of Reims Champagne-Ardenne, France
- **Fuchs, Christoph**, Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Germany
- **Gluhak, Alexander**, Intel Labs Europe
- **Hodoň, Michal**, University of Žilina, Slovakia
- **Johnsen, Frank Trethan**, Norwegian Defence Research Establishment (FFI), Norway
- **Krco, Srdjan**, DunavNET
- **Lenk, Peter**, NATO Communications and Information Agency, Other
- **Li, Guofu**, University of Shanghai for Science and Technology, China
- **Marks, Michał**, NASK - Research and Academic Computer Network, Poland
- **MURAWSKI, Krzysztof**, Military University of Technology, Poland
- **Niewiadomska-Szynkiewicz, Ewa**, Research and Academic Computer Network (NASK), Institute of Control and Computation Engineering, Warsaw University of Technology
- **Paprzycki, Marcin**, Systems Research Institute Polish Academy of Sciences, Poland
- **Sikora, Andrzej**, Research and Academic Computer Network (NASK)
- **Skarmeta, Antonio**, University of Murcia
- **Suri, Niranjana**, Institute of Human and Machine Cognition
- **Te Paske, Bert Jan**, TNO Netherlands
- **Wrona, Konrad**, NATO Communications and Information Agency
- **Xu, Jian**, Northeastern University, China
- **Zhang, Tengfei**, Nanjing University of Post and Telecommunication, China
- **Zhao, Yongbin**, Shijiazhuang Tiedao University, China

Adaptive Lighting System as a Smart Urban Object

Michael Aleithe
Leipzig University, Germany
Email: aleithe@wifa.uni-leipzig.de

Philipp Skowron
Leipzig University, Germany
Email: skowron@wifa.uni-leipzig.de

Eric Schöne
Eric Schöne Chartered Surveyor,
Germany
Email: es@schoene.co

Bogdan Franczyk
Wroclaw University of Economics,
Poland
Email:
bogdan.franczyk@ue.wroc.pl

Abstract—In this article we present an approach to an adaptive lighting system as an intelligent object supporting urban space, especially for the elderly. This intelligent lighting system is used as an instrument to improve the feeling of safety in everyday life by overcoming barriers such as dark areas at night. The intelligence of this system is based on a personalized and position-dependent adaptation of light, whereby intensity and color can be varied. This article focuses on the technical implementation of a corresponding lighting system. In this context, the main point of emphasis is the overall architecture, especially from the point of view of an application system.

I. INTRODUCTION

With regard to the digitalization of the urban environment, commonly referred to as Smart City, recent efforts in this area are now focusing on an increasingly ageing population. The latest findings will be described in [1] in this context. So-called smart urban objects can be used in public areas to increase the sense of security of older people, which can enhance their participation in public life. In [1] an information radiator is presented as a typical member of a smart urban object whose task is to boost the feeling of well-being by means of a tailored information supply for elderly people.

As another typical representative of a smart urban object, this work realizes a lighting system which aims to increase the feeling of security in this environment by a personalized adaptation of the light. The application portfolio of this adaptive lighting system is aimed at the personalized routing of elderly people, especially in rural areas, since the light pollution component is initially negligible due to too many people on one lighting section. The focus of this work is on the conception of a complete technical architecture as well as the description of the individual sub-components, which are necessary in the course of an implementation. So we can formulate the following research question: *How must an application architecture be designed in order to implement personalized and position-dependent adaptive lighting?*

The aspect of personalization is realized by a personal transmitter, which can be comfortably carried along by the elderly and can be used for distance measurement. Based on this personalized distance information, the light can be adapted to the intensity and color of the light. The psychological component of this light adaptation is excluded from the investigations in this article.

Section 2 of this article explains the preliminary work to date in the field of intelligent lighting systems. Subsequently, Section 3 describes the general technical system design with all associated subsystems for the implementation of a personalized adaptive lighting system. Based on this, a detailed examination of the specific elements of the subsystems required for an implementation is carried out in Section 4. Section 5 explains the advantages and disadvantages of the system, which can be evaluated based on the implemented prototype. In summary, Section 6 concludes with a final review of the adaptive lighting system.

II. RELATED WORK

In the past, various attempts tried to adjust the effect of light on adequate situations of everyday life in a target-oriented way. In [2], lighting in buildings was optimized so that sensors first register the existing natural light and, based on the acquired information, afterwards reduce the energy-bound electrical lighting to a minimum.

The minimized energy consumption also reduced the energy costs. This approach used wireless sensor networks in order to detect the already existing daylight. In contrast to the wire-bound variant, the wireless option showed various advantages. On the one hand, higher flexibility is given which makes the installation of the option much easier in already existing buildings. Further, the size and costs of the metering sensor system are being reduced due to the wireless variant. The network consists of an array of light sensors which communicate with a master node. That way, information about the current lighting state of the daylight are provided. The controlling of the lamps happens based on the master node.

A polychromatic lighting system, which is variable in terms of light color and intensity of lighting, was introduced by [3]. This way, lighting can be adjusted adequately to the needs of the user. Based on the already existing lighting intensity and light color, the realization of the light constellation is being regulated via linear and non-linear optimization techniques so that only a minimum of energy is needed. In order to apply optimization strategies in this case, the overall system had to be modelled. This includes measuring elements and lighting elements. The lighting elements include the already existing daylight, light bulbs, fluorescent lamps as well as LED arrays, though only the latter can be regulated in terms of lighting intensity and light color. The metering is done by sensor nodes and includes the recording of the light temperature of all light elements and the transmission of the information to the Controller. Optimization calculations are being done on the Controller in order to create a base for the regulation of the LED arrays and consequently the desired lighting constellation.

Based on the trend of using the already existing daylight more efficiently and thus to economize more ecologically and sustainably, [4] justify the so called connected lighting as an essential component for Smart Cities. The transfer from a traditional to an intelligent lighting was described. It was pointed out that the intelligence of this transfer is due to the interlinking of the individual lighting elements and thus the internal communication. The substantial advantages of the intelligent lighting in contrast to the traditional variant are remote monitoring, the technical components, a smart asset management as well as a constantly optimized reduction of lighting energy in order to economize more energy-efficiently. The result is a reduction of the lighting intensity of lamps on much frequented roads during quiet phases. In this source, the lighting concept which is connected to the internet is referred to as "Energy-Internet."

Smart lighting was used by [5] as a component to support sustainability and to improve quality of life of the participating populations in the Smart City context. Thus, a lighting system, which controls the street lights energy-efficiently without affecting the traffic situation or the social components of the area in a negative way, was presented. Smart lighting aims at operating adaptively and autonomously. The system includes intelligent controlling of street lights, which is based on adaptive behavioral rules. The traffic situation of the concerned area was registered via wireless sensor networks. This strategy was evaluated through an agent-based simulation. The result was a reduction of the spent lighting energy by 33%.

[6] investigated the realization of the light source adaptation. Focus of the investigation, in this case, was on the minimization of the energy consumption. The lighting intensity was adapted via a regulation – nominal value and real value of the lighting intensity are being compared continuously and in case of a discrepancy, the real value is being adjusted. The logic of rules is derived from fuzzy techniques.

[7] conducted a study in which various strategies for adaptive lighting were put into practice, the resulting energy gain was then analyzed. The portfolio of these strate-

gies includes, among other things, aspects of adaptive lighting. In order to realize an adaption to the environment, a sensor-based variation of the lighting intensity takes place. Further, an optimization concerning peak phase and low phase of the traffic to be handled takes place. During the peak phase, the lighting intensity increases and decreases during the low phase. Moreover, the area, which is analyzed in this study, has to be divided into zones and further inspected separately. The results of this study showed a significant reduction of energy costs, partly due to the strategies already introduced.

Another concept, introduced by [8], addresses the energy consumption of exterior lighting in cities. The concept includes a controlling system which, besides a minimized energy consumption, requests a remote diagnosis. That way, additional maintenance costs are expected to be avoided. The applied strategies are orientated towards the effects of the building automation on the building efficiency, as described in EN 15232 in more detail.

Ultimately, an immense saving of energy was proved. Concerning the minimization of the energy consumption of the used lighting components, [9] presented a method which optimizes the lighting intensity through an iterative method. In this context, a specifically centralized lighting control system was designed for this purpose. [10] reached a minimization of the illumination level through the maintenance of specifically defined nominal value and a steady minimization of the electrical energy. A cognitive approach through an adequate regulation algorithm was taken, whereby a non-linear characteristic of the illumination level was taken into consideration. Later, this approach was evaluated experimentally through [11]. In the end, an immense importance for a possible practical use could be proved.

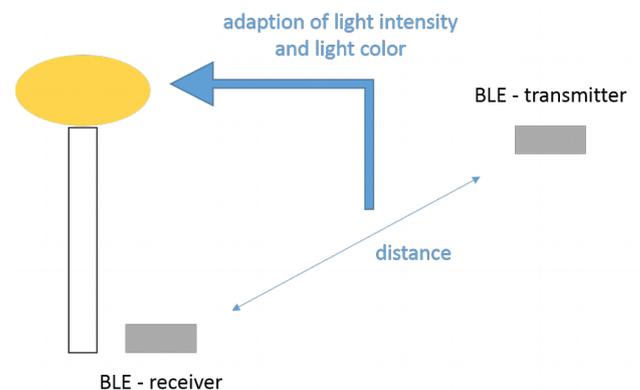


Fig 1. This picture illustrates the schematic overview of the system design in general. As consequence of the distance variability between receiver and transmitter and the coherent variation of BLE signal strength the parameter of the light behavior will be adapted. In this work we focused on intensity and color as light parameters.

[12] introduces a completely remote-controlled street light island as application in the Smart City context. The architecture with its diverse hierarchical structures is being described and the technologies used are being discussed in

detail. This includes mainly the controlling via Raspberry-Pi and wireless transmission via ZigBee.

In contrast to previous approaches, mainly explaining technical and to some extend economical concerns, [13] and [14] analyze the psychological effect of the lighting system on elderly people. Especially [13] aim at improving the mood of elderly people via adaptive lighting systems. Requirements for a technical system were defined, which detects the mood of elderly people and thereupon adapts the lighting in the room, aiming at activating the person by designing a specific mood-model.

III. APPROACH FOR SYSTEM DESIGN

Based on the research of previous work on intelligent lighting systems (Section 2), it becomes apparent that there is no system for personalized position measurement with coupled adaptation of lighting parameters in public spaces. The system architecture described in this section addresses this particular challenge. In principle, personalized distance measurement is performed using the Bluetooth Low Energy (BLE) standard [15]. A BLE transmitter is assigned to each person. The distance measurement is carried out by assigning a distance-dependent signal strength to the BLE receiver. This signal is called Received Signal Strength Indicator (RSSI). Starting from the incoming RSSI at the BLE receiver, the intensity and/or color of the light is adapted. This general system structure is illustrated in Fig. 1 for better understanding.

IV. IMPLEMENTATION

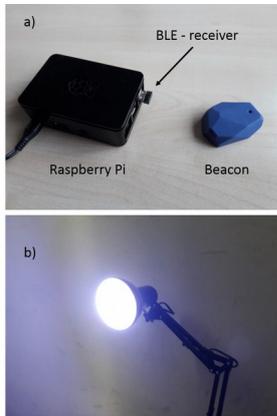


Fig 2. This illustration demonstrates the individual hardware components of the adaptive lighting system. In a) the BLE transmitter and receiver are represented, in b) the prototypical lamp fitting and in c) the lamp body for public space.

An Estimote Beacon is used as a personalized BLE transmitter [16]. The receiver is realized by a Bluetooth 4.0 adapter [17], which is able to receive the BLE standard. The transmission protocol is the so-called Eddystone protocol in UID format [18]. These received data (distance measurement) can then be further processed in a control unit and subsequently the adaptation of the lamps can be realized. Specifically, the lamp is controlled by means of a ZigBee

gateway, which can be addressed from the control unit via REST. The Philips Hue lamps [19] adapt their lighting parameters to the received control data of the ZigBee protocol. The BLE receiver, the control unit and the ZigBee gateway for lamp control are implemented on a Raspberry Pi. There is also a MQTT client on top of it. This serves as a further component for integration into a network and for monitoring purposes. The overall architecture with its individual modules described above is shown schematically in Fig. 3.

The implementation from the hardware perspective is illustrated in Fig. 2. a) contains the comfortable Estimote Beacon, which is used as a personalized BLE transmitter, and the Raspberry Pi with all the subsystems described above. Furthermore, b) shows the prototypical outfit of Philips Hue lamps as used for test purposes. In public space, these lamps and the Raspberry Pi are integrated into a lamp body as shown in c) [20].

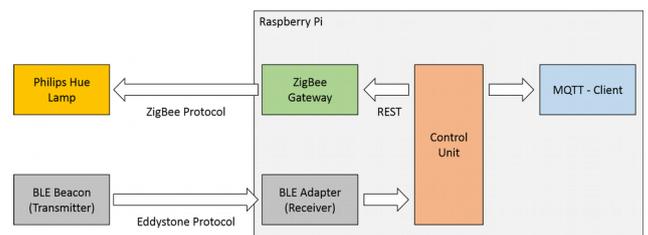


Fig 3. This picture demonstrates the architectural model of the adaptive lighting system with submodules and their communication dependencies.

V. RESULTS AND DISCUSSION

The system design described in Section 3, which has been discussed in more detail in Section 4 with regard to its technical implementation, is here subjected to an analysis of its advantages and disadvantages.

The implemented prototype is easy to use in several ways. On the one hand, the user does not need to operate the beacon (BLE transmitter) any further. On the other hand, the Beacon can easily be carried along in everyday life. In addition to this ease of use, Beacons generally have a long battery life, which further enhances the usability. The hardware required for the control (Raspberry Pi with subsystems) can be easily integrated into the lamp carcass, which is intended for later use in public spaces as illustrated in Fig. 2 c). Furthermore, a certain degree of flexibility can be observed from a software point of view, since standardized protocols and formats are used throughout (Section 4), so that individual subsystems are relatively easy to replace.

Furthermore, the quality of the BLE signal was examined in detail. In this context, the RSSI was measured depending on the distance between the transmitter and receiver. The measurements were carried out both with and without a lamp body to determine the influence of the additional object on the signal absorption. For each constellation 10 measuring points were recorded. The results are shown in a Box-Plot diagram in Fig. 4. Here, a relatively high variability

ity of the measured values can be observed, which means that an exact distance measurement is not necessary. The causes of these fluctuations can have many different causes, since BLE reacts sensitively to liquids, WLAN and metallic bodies [21, 22, 23], but these cannot be prevented in public space. In general, a signal attenuation through the lamp body can be determined by the measured values. In both cases (with and without body), however, it is possible to differentiate according to the measured values of the RSSI whether a beacon is located inside or outside a radius of about 5 meters around the receiver.

VI. CONCLUSION AND OUTLOOK

In this work, a novel smart urban object is presented, which is realized by a personalized adaptive lighting system. The focus is on the technological architecture. A concept for an overall architecture and all subsystems involved was presented, which were necessary for the implementation. The completed prototype was briefly presented and evaluated according to various criteria. In summary, it is a personalized adaptive lighting system that provides a distance-dependent variation of the intensity and/or color of the light, whereby the accuracy of the distance measurements is limited to a differentiation between greater or less than 5 meters.

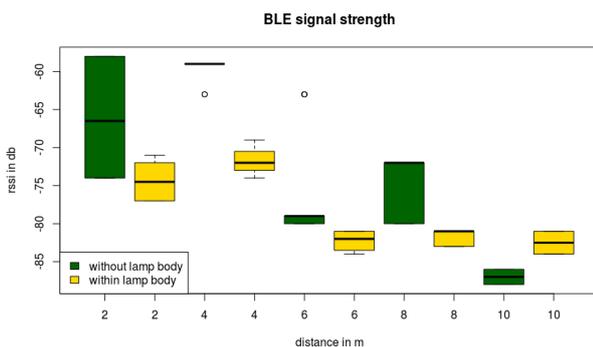


Fig 4. In this boxplot-diagram we illustrate the numeric relation between the Received Signal Strength Indicator (RSSI) of BLE signal and the distance of the BLE -transmitter and -emitter. The measurements conducted with the BLE-sensor without (green) and within (yellow) the lamp body to investigate the influence of the lamp body.

Due to this measurement disturbance, data-technical corrections which increase the precision of measurement will be offered in the future. Also, this lighting system will be installed in a test site that ideally reflects the public space. For this reason, an investigation of the psychological aspects is necessary to what extent adaptive light influences the mood of the participating persons. Furthermore, the realization of a complex network of adaptive lighting stations is aimed at, so that entire areas of the public space can be covered. In addition, the existing lighting system must be expanded to include several people or groups of people.

ACKNOWLEDGMENT

This work was fully conducted in the scope of the research project *UrbanLife+* (16SV7442), funded by the German Federal Ministry of Education and Research.

REFERENCES

- [1] Koch, M., Kötteritzsch, A., Fietkau, J.: Information radiators: using large screens and small devices to support awareness in urban space. Proceedings of the International Conference on Web Intelligence (WI '17), pp. 1080–1084. ACM, Leipzig (2017)
- [2] Nippun Kumar, A. A., Kiran, G., Sudarshan, T. S. B.: Intelligent Lighting System Using Wireless Sensor Networks. International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) 1 (4), pp. 17–27 (2010). doi:10.5121/ijasuc.2010.1402
- [3] Paradiso, J. A., Aldrich, M., Zhao, N.: Energy-efficient control of solid-state lighting. SPIE Newsroom. (2011). doi:10.1117/2.1201102.003543
- [4] Crowther, J., Herzig, C., Feller, G.: The Time Is Right for Connected Public Lighting Within Smart Cities. Cisco Internet Business Solutions Group (IBSG) (2012)
- [5] Escolar, S., Carretero, J., Marinescu, M., Chessa, S.: Estimating Energy Savings in Smart Street Lighting by Using an Adaptive Control System. International Journal of Distributed Sensor Networks, pp. 1–17. (2014). doi:10.1155/2014/971587
- [6] Papantoniou, S., Kolokotsa, D., Kalaitzakis, K., Cesarini, D. N., Cubi, E., Cristalli, C.: Adaptive lighting controllers using smart sensors. International Journal of Sustainable Energy 35 (6), pp. 537–553 (2014). doi:10.1080/14786451.2014.923887
- [7] Jackson, H., Jackson, S., Jackson, C., Siminovitch, M.: Saving Energy in Buildings with Adaptive Lighting Solutions. California Energy Commission. California Lighting Technology Center, UC Davis (2015)
- [8] Ozadowicz, A., Grela, J.: Energy saving in the street lighting control system—a new approach based on the EN-15232 standard. Energy Efficiency (2016). doi:10.1007/s12053-016-9476-1
- [9] Caicedo, D., Pandharipande, A.: Daylight and occupancy adaptive lighting control system: An iterative optimization approach. Lighting Research and Technology 48 (6), pp. 661–675 (2016). doi:10.1177/1477153515587148
- [10] Yin, C., Stark, B., Chen, Y., Zhong, S.: Adaptive minimum energy cognitive lighting control: Integer order vs fractional order strategies in sliding mode based extremum seeking. Mechatronics 23 (7), pp. 863–872 (2013). doi:10.1016/j.mechatronics.2013.09.004
- [11] Yin, C., Stark, B., Chen, Y., Zhong, S., Lau, E.: Fractional-order adaptive minimum energy cognitive lighting control strategy for the hybrid lighting system. Energy and Buildings 87, pp. 176–184 (2015). doi:10.1016/j.enbuild.2014.11.036
- [12] Leccese, F., Cagnetti, M., Trinca, D.: A smart city application: a fully controlled street lighting isle based on Raspberry-Pi card, a ZigBee sensor network and WiMAX. Sensors 14 (12), pp. 24408–24424 (2014). doi:10.3390/s141224408
- [13] Huldgren, A., Katsimerou, C., Kuijsters, A., Redi, J. A., Heynderickx, I. E. J.: Design Considerations for Adaptive Lighting to Improve Seniors' Mood. Inclusive Smart Cities and e-Health, Bd. 9102. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 15–26 (2015)
- [14] Kuijsters, A., Redi, J., Ruyter, B. d., Heynderickx, I.: Lighting to Make You Feel Better: Improving the Mood of Elderly People with Affective Ambiances. PloS one 10 (7) (2015). doi:10.1371/journal.pone.0132732
- [15] Specifications: The building blocks of all Bluetooth devices. <https://www.bluetooth.com/specifications>. Accessed on 3 Jan 2018
- [16] The Physical World: Software-defined. <https://estimote.com/>. Accessed on 3 Jan 2018
- [17] Bluetooth 4.0 USB Adapter. <https://www.asus.com/Networking/USB400/>. Accessed on 3 Jan 2018
- [18] Eddystone-UID. <https://github.com/google/eddystone/tree/master/eddystone-uid>. Accessed on 3 Jan 2018
- [19] White and colour ambience. <https://www.philips.co.uk/c-p/8718696461679/hue-white-and-colour-ambience>. Accessed on 3 Jan 2018
- [20] https://www.burri.shop/resources/Public_Light_small_reflektor.jpg. Accessed on 3 Jan 2018
- [21] Estimote Beacons: Will wireless interference and Wi-Fi impact beacons? <https://community.estimote.com/hc/en-us/articles/200794267-Will-wireless-interference-and-Wi-Fi-impact-beacons->. Accessed on 3 Jan 2018

- [22] Estimote Beacons: Best practices for installing Estimote Beacons. <https://community.estimote.com/hc/en-us/articles/202041266-Best-practices-for-installing-Estimote-Beacons>. Accessed on 3 Jan 2018
- [23] Estimote Beacons: What are Broadcasting Power, RSSI and other characteristics of a beacon's signal? <https://community.estimote.com/hc/en-us/articles/201636913-What-are-Broadcasting-Power-RSSI-and-other-characteristics-of-a-beacon-s-signal->. Accessed on 3 Jan 2018

Real Time Risk Monitoring in Fine-art with IoT Technology

Mark Phillip Loria, Marco Toja

See Your Box
2 Cormont Road
London, England
SE5 9RA (UK)

Email: {mloria, mtoja}@seeyourbox.com

Vincenza Carchiolo, Michele Malgeri

Università di Catania,
Dip. Ingegneria Elettrica Elettronica e Informatica,
Viale Andrea Doria 6,
95125 Catania, Italy

Email: {vincenza.carchiolo, michele.malgeri}@dieei.unict.it

Abstract—This work presents a bespoke system used to monitor inter-modal logistics within the fine arts industry. A custom IoT architecture provides end-to-end capabilities allowing continuous risk assessment during storage, handling, transport and exhibition. The system overcomes the challenges of implementing adaptive artificial intelligence systems, extra low latency and exceptional power efficiency within a fully integrated IoT architecture. The main contribution of this paper lies in the architecture that has been fully implemented and commercialized to international leading companies.

I. INTRODUCTION

EVER since informatics exist, there has been a great interest in the cultural heritage application, many projects aim at improving artwork usability and preservation exploiting some ICT technology and today it is one of the driving forces for both preserving and exploiting Cultural Heritage. Of course, changes in technology and market conditions created and continued to create new and more ambitious challenges [1]. ICT gives rise to a rapid and substantial change in the practice of utilization, supply, and conservation of cultural heritage [2].

Today, virtual reality and IoT are the breaking technologies in cultural heritage and art management. Both technologies play an important role in enhancing the user experience inside real museums, they allow to create complete virtual museums or personalize the surrounding environment. The cultural heritage is changing thanks to the resources offered by ICT technologies. The impact of ICT spans over different application and examples from management to new ways access to arts and culture.

The years to come are going to be interesting and challenging for the use of Internet of Things (IoT) in managing artwork. In this context, some solutions adopted in the supply chain [3] are the starting point for a model to monitor transportation. Indeed, See Your Box™ Ltd., since its foundation, uses IoT technologies in several applications for arising problems close to artwork management [4], [5].

IoT is one of the ICT technologies that has the largest economic impact on art and culture in many areas such as

medicine, energy management and many others. Of course, also in the field of cultural heritage, the IoT is becoming a fundamental technology since it allows the development of several applications covering different objectives. In the last years, two of the main areas in which IoT leads are usability and monitoring.

This work presents a bespoke system with a customized architecture to monitor the transportation of works of art. The system presented in this work has been already implemented and released on the market.

Section II discusses some related works about the use of IoT in the cultural heritage framework. Section III shows the proposed architecture for creating a data collection system for a real time risk monitoring. The main contribution of this paper is discussed in section IV where the architectural solution tailored to inter-modal art logistics processes is presented. Moreover this section discusses the solution proposed by See Your Box™ Ltd.:

Finally, section V highlights some conclusive remarks with particular attention to the impact of the product on the market.

II. RELATED WORK

The most consolidated applications based on the IoT paradigm aim to improve user interactions with museums. There are numerous examples using IoT to craft smart museums [6]. These applications enhance the user experience guaranteeing better interactions with the artworks and their history often realizing smart virtual environments. The main goal of these applications is to renovate the users' interest on the cultural heritage, by enhancing the cultural experiences. Usually, these applications use an indoor location-aware architecture able to provide users with more information and media experience either inside a museum [7] or in open spaces. The most common enabling technologies are (nearly) passive systems such as RFID, Bluetooth, NCS [8][9].

One of the most common examples is the development of a smart guide that helps visitors learn more about artworks in the museum. This application is often enhanced by an artificial intelligence able to recognize speech and/or perform some image processing in order to provide a smart guide. Other interesting examples aim at monitoring the visitors' behaviour in order to

This work was funded in part by University of Catania, Dip. Ingegneria Elettrica Elettronica Informatica (DIEEI), under DEDuCE project

improve the usability of the museums. For example, Cuomo in [10] uses IoT to collect data about visitor's behaviour and clustering techniques to classify spectators in order to reproduce the visitors dynamics using statistical method.

As previously mentioned, IoT paradigm finds an application in the realization of fine art monitoring systems. Monitoring can be planned in-situ and/or during transportation. Of course, the involved IoT technologies must be tailored to the different use-cases. Therefore the selection of proper IoT technology to monitor the artwork in-situ could be different from the one used to monitor the artwork transportation. The former are IoT indoor technology (e.g. RFID) while the latter create new challenges and interest [11][12]. Usually these applications share the common problem of collecting some sets of values such as, temperature, humidity, vibrations, etc. that can be used to evaluate whether the artwork has been properly managed in-situ or during transportation (e.g. remembering any over-temperature and its length). Moreover, the IoT technologies involved in monitoring during transport must be able to collect also location information. The problems connected to this type of application stem mainly from the sensitive nature of the artworks themselves, because they are fragile against different factors, like vibrations, shocks, temperature, humidity, etc.

A relevant project in monitoring of transportation is PACT-ART [11][13] helps smarten the way transportation processes using external and IoT services to handle tasks depending on their contexts. The system draws reliable predictions about potential misbehaviour and yet-to-come violations.

However the existing solutions often analyze the collected data only at the end of transportation thus they are able only to signal eventual anomalies. Indeed, they cannot perform any action that can try minimize the impact of the anomaly so reducing eventual damages. We believe that the capability to avoid or reduce the damage is the main matter both in economic terms and in cultural and historical terms, than the solution proposed in this paper.

III. MOTIVATION

In this section we discuss the proposed architecture for creating a data collection system for a real time risk monitoring solution tailored to inter-modal art logistics processes. Monitoring risk in a inter-modal logistics processes represents a number of challenges for the ICT solutions involved. Across the process multiple stakeholders are concerned about very different KPIs while goods are handled in different ways and challenges can be of different nature. Furthermore, as it may sound, the actual shared goods across the process are not the paintings but risk. Risk for the collector is the health of the paint on an ancient unique painting, risk for the logistics company is tied to the handling of an insulated shock absorbing crates, for a museum it's temperature control inside a safe and so on.

The further challenge for the data collecting solution is that risk is not necessarily (only) directly related to plain data. A very good example is described in [11]. The integrity of an art piece depends on temperature. Supposing the painting is

transported in an insulated case, a measured value approaching critical levels might be absolutely fine if the external temperature exhibits an opposite trend. On the contrary, small variations measured inside the crate could be an important indicator of either a damaged box or a dangerous condition. In other words the temperature level measured close to the painting is not sufficient to determine the level of risk without knowing the temperature outside the box, the time of year and day, the latitude, weather forecast and how long the painting is expected to stay in that condition. Data without specific context in an inter-modal monitoring provides data that most of the time are either incomplete or irrelevant. In literature it's possible to find many references on work performed in the field of context aware data analytics systems. These make it possible to identify the different phases of an inter-modal logistic process and adapt the analytics and business intelligence accordingly. As an example, taking into consideration data collected from an accelerometer sensor, the same data set could have two completely different meanings if collected during two different steps of the logistics process. In figure 1 samples collected at 100 Hz from a high precision MEMS sensor during transport reveal a perfectly safe condition since the resulting G force is always below the maximum allowed threshold of 3G for fine art. However, when the same data is sampled in a stationary context 2 the system is able to detect a potential threat since the art piece is supposed to be absolutely still during an exhibition.

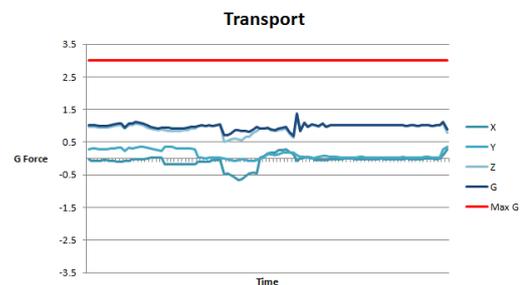


Fig. 1. Accelerometer profile during transport context

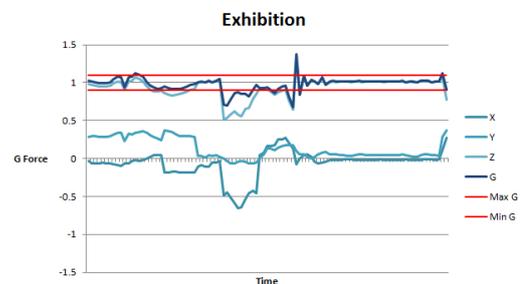


Fig. 2. Accelerometer profile during exhibition context

It's also possible to find in literature references on context-aware monitoring solutions targeted to the fine art logistics

industry and the focus on this specific field is particularly relevant due to:

- The economic and cultural value of the moved goods
- The goods are typically unique and not replaceable
- The high insurance premiums for theft and mishandling
- The phases of the logistics process are standard and well planned

However, while tackling sharply the matter from a data analytics view point, many solutions overlook how to source data of moving art pieces. In this paper we want to shift the focus from data analytics and information extraction to the data collection phase. Data analytics is surely a great tool for extracting a truthful risk model that can lead to better understanding of the underlying patterns. On the other hand, due to the value and uniqueness of fine art pieces stakeholders are demanding early or real time access to a measure of risk profile for the art pieces while they are moved across the globe on an inter-modal process. This step represents one of the most challenging components of a real time monitoring system for fine art logistics, especially in the case of temporary exhibitions. This implies that any data collection system must be able to record and transmit data globally without the need of a dedicated telecommunication infrastructure. The system must be battery powered, wireless and its dimensions must be such to allow a pervasive monitoring of the art piece. These points lead to a real time monitoring system that must be adaptive from its core by adapting not only the analytics but also how data is collected. The system must also be able to distinguish among different masterpiece classes. Paintings differ in material and era, and are not to be treated as sculptures or jewelry. Inter-modal risk monitoring requires therefore an embedded AI system capable of adapting both to a variety of goods and a variety of logistics scenarios. If not, such lack would force to create so many different technologies that would result in an unsustainable business model.

Following the previous example, the effects of lowering the accelerometer sampling rate to reduce the amount of data transmission in order to contain battery drain and telecommunication costs, could have potentially dangerous effects. In figure 4 we show how lowering the sampling rate effects the resulting data acquisition and subsequent analysis.

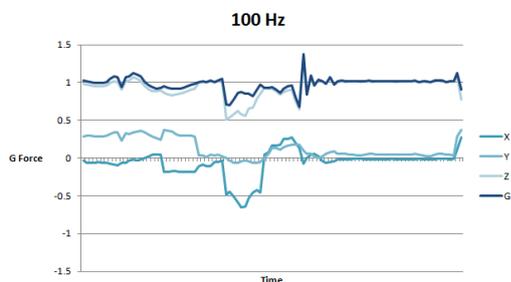


Fig. 3. Accelerometer profile sampled at 100 Hz

During transport on road low intensity high frequency vibrations caused by rough tarmac or cobblestone can have

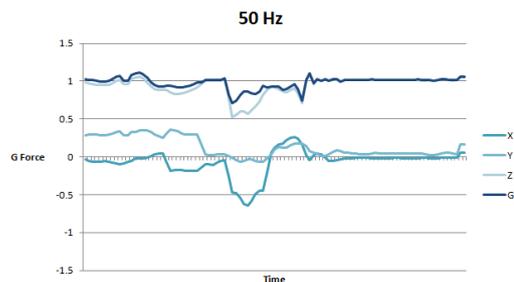


Fig. 4. Accelerometer profile sampled at 50 Hz

devastating effects on moving art pieces. When a design setting such as accelerometer sampling rate is fixed to a low value to contain battery drain during the months of stationary exhibition it could totally miss capturing potentially dangerous situations that could occur during the few hours of road transport. On the other hand configuring the system for a worst case scenario would have negative effects on battery drain, transmission costs and memory usage during periods of signal loss when data is buffered. The challenges encountered have already been a topic of research and we identified two macro categories of approaches:

- Centralized monitoring - Data is collected by simple sensors that have no logic applied. This approach is typical of wireless sensor networks. A central node with high computational power processes incoming data to detect an event
- Distributed monitoring- This approach is typical of edge computing or smart sensors. These have enough processing power and a sufficient vision of the environmental conditions to detect events.

Both approaches have their limitations when confronted with the challenges of fine arts. In a centralized approach customization of sensors is done at design time. Elements such as sampling frequency, resolution and power management are set for the whole process leaving very little room for optimization. Sensors are very cheap due to the low complexity but require continuous stream of data to the centralized node. By leveraging edge computing, distributed systems overcome the issue of data transmissions at the cost of complex and more expensive sensors. These however are often lacking a complete overview of the process since context awareness is achieved by the interaction with other sensors and require high computational power in devices that need to be almost disposable and therefore are cost sensitive. These considerations lead us to the architecture presented in this paper, that is a fusion of both approaches. A high performance cloud based central node continuously segments the risk model into sub-models for each one of the inter-modal logistics phases translating them into simple configuration parameters for the sensing nodes. When the system detects the occurrence of an event or change in context the parameters are uploaded to the distributed nodes. The rationale behind the architecture is quite

simple, since only a node with full vision over the process can determine what risk sub model fits the process at the give time. Understanding these restrains and specific legal requirements leads to build a dedicated hardware and software architecture where the key elements are the speed, size and content of the package you transmit.

IV. ARCHITECTURE

The system is composed by two main elements that are the smart sensing devices and a centralized cloud based infrastructure with a high performance rules engine that processes incoming data. Due to the volume of data the whole cloud platform is built upon a totally virtualized environment to allow instantaneous scalability. A web-based management dashboard offers real time access to the risk monitoring platform while REST APIs enable full M2M integration with the system. The architecture of the cloud platform has been described in detail in previous work. In the rest of this chapter we will discuss in detail the architecture of the smart sensing device.

A. Smart sensing devices

The sensor nodes are based on a proprietary hardware platform. One of the key ideas of IoT for industrial applications is the possibility of digitalizing a process by capturing data points directly on goods. A concept built on the experience of the flight simulators for military applications that is seeing new usages in the industrial world, is the concept of digital twin. Data collected on the real object feeds a computerized model that is used for predictive analysis and simulation. To achieve the necessary level of confidence it's mandatory to have sensors that can be placed as close as possible to the real object and capture key parameters. In the case of fine art paintings this means placing the smart sensing devices directly on the picture frame as close as possible to the canvas and not on the building or vehicle. In figure 5 a smart sensing device is placed directly on a picture frame during the packaging phase and will be removed only after the end of the whole logistics process six months later.

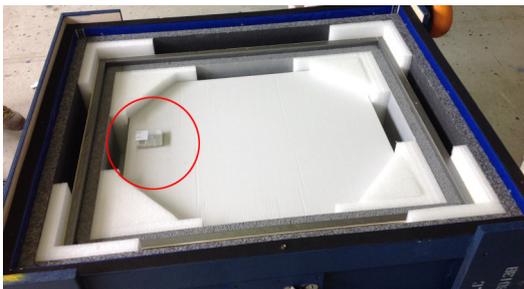


Fig. 5. A SYB device placed directly on a painting frame before placing inside transport box

This creates a number of challenges, specially for end-to-end monitoring. From a hardware and mechanical view point the smart sensing devices will have strict constraints on size, weight and type of batteries. From a usage process view point the device must be completely autonomous after power on and

require no human intervention until completion of monitoring. From a software view point the smart sensing device must be able to collect and process data as specified by the risk sub-model, guarantee time synchronization of collected data, manage power and data transmission efficiently and lastly ensure a high level of data protection for transmitted and stored data. Figure 6 summarizes the functional blocks of the smart sensing device and in the rest of this paragraph we will go into detail for the most important ones.

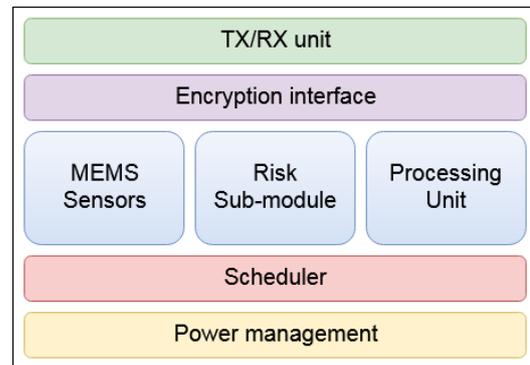


Fig. 6. Functional blocks of the smart sensing device

1) *Risk sub-model parameter set:* In the hybrid architecture we are proposing elements of the risk model are shared with the sensing nodes. Transferring a whole risk model, that can easily be as large as 2 MB, to the edge nodes would raise the minimum requirements in terms of memory, processing power and transmission costs. On the other hand these devices must be as simple and low cost as possible since in many situations they might need to be considered as disposable. The underlying idea is to first break up the risk model into sub-models, one for each phase of the process, and subsequently translate them into a small set of incremental parameter changes for the smart sensing device. As an example, accelerometer impact profiles and vibrations measured over more than 100.000 working hours of monitoring service were divided into risk classes. Using data mining techniques for each class we were able to extract key parameters for the processing unit that detects potential impacts such as maximum G force during impact, minimum sampling frequency, low-pass and high pass filters, vibration waveform and so on. To further optimize the transfer of these parameters we restricted them to a discrete amount of reasonable values and structured them inside configuration registers. To put this into perspective with a practical example, when the transport of a painting is completed and it is placed into the museum the change of parameters is performed with the transmission of only 8 bits of data. This is sufficient to instruct the smart sensing device to change the type of accelerometer event to detect and adjust the thresholds accordingly. It is important to underline that this simplification of the risk sub model affects only the smart sensing nodes and data incoming to the cloud platform is processed using the full risk sub-model. Of course when using a discrete amount of parameters it is only possible to customize the functionalities

at run time according to what has been defined at design time. This however is not a limitation since the responsibility of the smart sensing device is not to determine if the actual event has happened (calculating therefore risk) but collecting relevant data for the centralized processing node. The collection of false positives actually is beneficial for the system to feed the data mining algorithm with more balanced data. The risk sub-model parameter set is not only in charge of determine how data is collected but also how data is transmitted by the smart sensing device. On this matter there are at least three crucial elements that it must manage. The first one is instructing the smart sensing device on contexts where data transmissions are not allowed, the second one is optimizing means of data transmission for costs and the third one is power consumption optimization. During end-to-end monitoring for temporary exhibitions it is not uncommon to have to perform a logistic step over air freight. Precise regulations (quote) specify how and when portable electronic devices are allowed to be active and transmit. In the case of fully autonomous devices such as the ones we are describing in this work, this creates a strong challenge to guarantee the compliance with existing regulations while the system relies totally on AI. To achieve a high level of reliability it was necessary to create a three level failsafe aircraft detection algorithm that uses data from three different sensors. All related algorithms have been coded in negative logic so in the event of failure on one sensor the smart sensing device will enter a protection mode and disable transmission not to risk transmitting while on an aircraft. Art pieces moving around the world and actively transmitting data requires careful attention on optimizing what channels to use to transmit data not to incur into heavy roaming charges. The risk sub model contains all necessary parameters to regulate what channels to use, for instance GSM rather than 3G data, to optimize not only cost but also power consumption.

2) *Scheduler*: In the previous section we went through an example regarding the customization of the accelerometer. The smart sensing device is equipped with 7 physical sensors and is able to interpolate them to create virtual sensors. An example is crash detection that is achieved with an interpolation of data from accelerometer and microphone. For each sensor the smart sensing device can set thresholds, sampling rate and pre processing data algorithms to execute. The internal scheduler acts as the heart of the smart sensing device, regulating precisely how data is sampled from sensors and generates all necessary tasks that in turn are executed. Lastly, and most importantly, it is in charge of keeping all data collected in sync with the time reference of the centralized node. The scheduler and hardware architecture are heavily coupled. During system design we identified two macro approaches that were either adopting a high performance micro controller with low cost simple sensors or a simpler processing unit and more complex MEMS sensor that could be programmed. The former was quickly discarded since it would have required almost for sure a real time OS and the designing of a scheduler with a high frequency internal rate to poll all sensor at an adequate rate. The latter, on the other hand allowed us

to design a simpler scheduler, run with a slower internal clock and ultimately lower power consumption. In the final implemented design all interactions with sensors are based on an asynchronous approach leveraging internal threshold based interrupts. This allows us to keep the main scheduling loop for task management to a very low frequency. When collecting data used to generate a risk profile it is absolutely crucial to guarantee perfect time synchronization and to know exactly when an event has actually happened. Smart sensing devices, or any remote node in general, will be subject to two challenges related to time synchronization of data. The first one is managing offset synchronization while the second one is managing imprecision on the oscillator of the devices itself. This is particularly relevant when designing an ultra low cost device that can collect data in hostile conditions since temperature can affect the precision of the oscillator. Common solutions, such as synchronization with external references like GSM networks or GPS satellites were not viable. Currently there is no international agreement on time synchronization between GSM providers worldwide and the availability of GPS coverage can be a serious issue for indoor monitoring such as museums and safes. To further challenge the system, to contain power consumption active components (and therefore clocks) are off or in an ultra low power mode during most of the time. The adopted strategy was to use two internal clocks (readily available in the chosen micro controller) to self calibrate the real time clock used by the scheduler. With this simple solution we were able to increase the accuracy of a factor of 10 while not introducing extra hardware. A second later of synchronization to manage potential drift issues is introduced on the central processing node. This layer uses a tick based approach where incoming time references from the smart sensing devices are compared with NTP servers and samples are repositioned in time using a linear projection. While introducing an element of complexity due to a dedicated time synchronization algorithm this approach allowed the system to achieve a higher level of accuracy in harsh conditions over alternative solutions such as one-time synchronization coupled with a CMOS backup battery. As an added benefit, it also allows to contain cost and space on the PCB.

3) *Power management*: The smart sensing device can be used for monitoring long term end-to-end exhibitions that can last up to six months. To allow the device to be placed directly onto the picture frame weight and dimensions of the batteries must be contained to a minimum. Therefore it's necessary to enforce strict power management techniques to minimize consumption. The peculiarity of end-to-end inter-modal logistics transports is that data can be collected and transmitted in very different ways. As an example, to ensure that paintings are transported onto the safest and fastest route it is crucial that location updates must happen every 10 minutes to allow the centralized node to detect potential problems like the presence of exceptional traffic and therefore allow a rerouting solution in time. During an exhibition on the other side daily data updates are sufficient. These are two examples of how it's possible to leverage the context of two

different steps within the logistics process to optimize power consumption. However, transmission rates are not the only area of possible intervention. Other examples are the need of having an RF interface always on to receive asynchronous requests from the central node. These are crucial during transport to allow the central node to ping the location of the smart sensing device at any time. It's not always possible or optimal to set these parameters within the risk sub-model and statically impose them to the scheduler. The reason is that many of these decisions require a large amount of environmental parameters and transmitting them to the centralized node would defeat the objective of reducing the number of transmissions and quantity of data. The key idea we are using is to insert within the risk sub model parameters that can define if a specific power saving strategy is allowed or not and let AI algorithms on the smart sensing device make the final decisions. The challenge of this approach however is that while it's possible to find many algorithms and tools with high levels of abstraction that are easy to implement and run on a server, scaling them down to an ultra low power micro controller with limited quantity of memory and resources requires custom solutions and proprietary algorithms.

4) *Data processing*: One of the benefits of not creating a purely telemetric system is that data can be processed on the edge node. This also allows us to heavily reduce the amount of data that is sent from the device to the centralized node during a standard end-to-end fine art logistic process that could easily last over six months. A telemetric system that monitors acceleration at 50 Hz would need to transmit, supposing 16 bit resolution for each axis, over 4.5 Gb of raw data streamed to the centralized node. From a cost perspective this absolutely not acceptable since part of inter-modal logistics are international and transmissions would incur into massive roaming charges. Even buffering chunks of 1 minute would always mean facing over 250.000 data transmissions. The battery would not last more than a day. On the other hand a system with edge computing allows us to implement a much more efficient strategy. If we are monitoring impacts we can transfer the maximum impact and type of event to detect within the risk sub-model using 16 bits and each acceleration peak would only require 6 bytes of data to transmit. Even if the risk sub-model on the server would require a full minute of data recording to analyze thoroughly the impact we would have reduced the amount of data to 18Kb for each potential impact event (16bits per axis sampled at 50Hz for a minute). The basic strategy is to use the risk sub-model parameters to reduce the amount of data by sending data with a higher content of information and less noise.

5) *Security*: The collection of data related to things is a sensitive topic as it raises many concerns regarding security. While security protocols and policies are not a new topic and we can rely on well established best practices for IP based solutions, when scaling them down to low power IoT devices they either result cumbersome or simply not relevant [14]. Designing an effective solution for IoT is still an open challenge [15]. In the fine art logistics industry the major

challenge is protecting data from malicious activities related to theft. It was reported that it took only 20 seconds to steal fine art jewellery from an Italian museum causing a total loss of several million [16]. Such efficiency is generally achievable only after a careful study of guard habits and security measures of a museum. Today smart sensing devices are used during exhibitions to monitor the passage of guards and ensure that the painting is safe. Should this data become available to a thief it would put the safety of the painting at great risk. To protect the system prior to a potential theft it is essential that the device and transmission protocols offer a form of protection against eavesdropping. All data exchanged between the smart sensing device and the centralized node is sent over encrypted channels that use a proprietary encryption algorithm where a symmetric encryption key is transferred over a dedicated channel. In the unfortunate event of theft the system must be able to face at least two other types of attacks. These are typically replay attacks and man-in-the-middle attacks. These types of attacks, if successful, could allow a thief to fool the system by making it believe that the painting is still located in the same position or even worse alter data altogether. In both cases risk is heavily mitigated by the use of encryption, nonces and hardware anti tampering mechanisms on the device itself.

B. Cloud platform

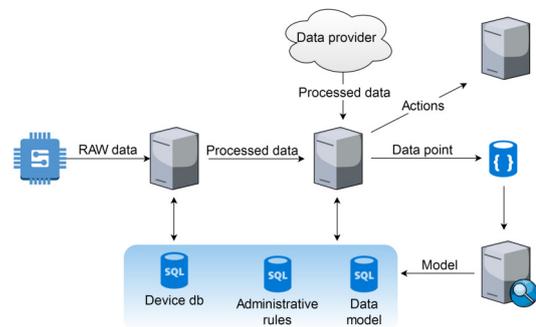


Fig. 7. See Your Box Architecture

The central processing node is based on a fully scalable infrastructure based on virtual machines that are responsible of fulfilling specific tasks. A lot of research and effort was invested in creating an efficient self load balancing system that could use the full potential of the available hardware. This allows the system to take advantage of instant and dynamic vertical scaling driven by the actual load of the system. The resulting architecture is summarized in figure 7. The cloud platform is subdivided in three main component:

- Gateway, accepts incoming requests from devices, authenticates and decrypts data, forwards data within the system and delivers messages to devices.
- Rules Engine, applies business logic to the incoming data according to the risk sub model

- Databases, the system uses a multiple DBMS approach to leverage the strengths of relational and non relational engines [4]

Details regarding the internal architecture of the system are beyond the scope of this article, however for completeness we will briefly consider in detail the element of task scheduling. The nature of communication channel used to transfer data from the smart sensing device to the cloud infrastructure is characterized by a very short timeout. The reference specification is that all incoming requests will timeout within a second.

The risk model that is used on the centralized node is structured as a set of rules that have a specific priority rank. All business logic is applied by executing these rules. Examples are data decryption, time synchronization or a change from a risk sub-model to another one altogether. Some of these rules are low latency (e.g. conversion of raw sensor data to a float value) while others are high latency due to complexity or interaction with external providers (e.g. compare sampled temperature with weather forecast service). Some rules have effects on the risk sub-model (e.g. a geofencing rule registers the arrival of a painting to the destination museum) while others don't (e.g. decrypt incoming data).

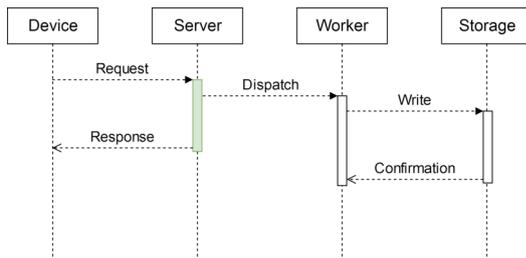


Fig. 8. A low priority high latency task is processed asynchronously

High priority tasks that are used to determine changes into the risk sub-model are placed into a synchronous queue so that effects of processing can be returned immediately to the smart sensing device. Low priority tasks are placed into an asynchronous queue so that we can close the communication channel as quickly as possible. Keeping the channel open for the minimum time necessary has a strong impact both on battery life for the smart sensing devices since all RF circuitry is active and on the performance of the cloud infrastructure since the gateway needs to be able to process massive quantities of requests per second.

C. Web based risk monitoring dashboard

The last component of our real time risk profile monitoring platform is the dashboard. It allows stakeholders involved in the inter-modal logistics process to have a real time over view of what is happening to the transported or stored painting. The dashboard was developed using AngularJS, a Javascript framework for single page applications and the layout of content was managed with the HTML/CSS framework Bootstrap. When using the dashboard users can:

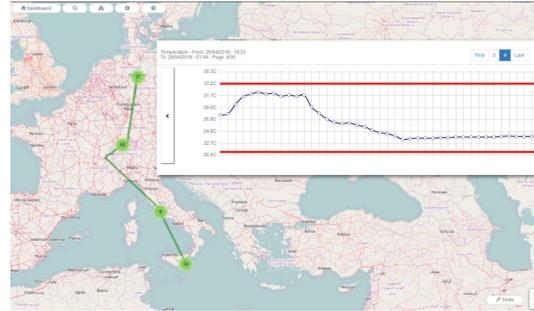


Fig. 9. Web based real time monitoring dashboard

- View the position of the painting over a map
- View time-series charts of all recorded environmental parameters
- Export data as CSV or PDF
- View the list of alerts generated during a monitoring activity

The dashboard was developed as a responsive web application that can be accessed also from mobile devices. Even though the true potential of the system lays in the M2M REST APIs that offer full integration with external sensors, the availability of a full fledged dashboard is essential in an inter-modal logistics process since multiple stakeholders might be part of different companies and a process of integration with each one would be slow.

D. A Case Study

The system presented in the previous chapter has been used to monitor over 200 real fine art international logistics transports and over 10 full inter-modal long term monitoring projects of temporary exhibitions that involved the coverage of both transport and exhibition. The logistics services were provided by a leading and specialized company for fine art transport while the exhibitions took place in museums amongst which Guggenheim and Vatican Museums. The duration of the long term monitoring projects, from nail-to-nail, ranged between 3 and 6 months. During these the system was able to swap into different configurations by accurately detecting the context of where the painting was placed. These swaps allowed the system to not only optimize power consumption, allowing six months of continuous real time monitoring while using a standard battery for mobile phones, but most importantly detect potential threats that were handled before they could actually damage the paintings.

As an example in figure 10 an inter-modal fine art logistics process over land and sea shows both the near miss and the trespassing of the critical temperature thresholds that the painting can withstand. Knowing the context the system was able to predict the risk for potential damage preventively during transport, allowing the logistics company to intervene in time and act upon the temperature control unit. In the second instance the system was able to ignore the potential threat since the combination of light levels, localization and movement

were sufficient to detect the detachment of the sensor from the picture frame and therefore the end of the monitoring activity.

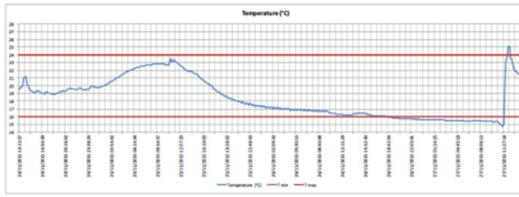


Fig. 10. Near miss of temperature thresholds

V. CONCLUSIONS

The paper discussed a systems, developed by See Your Box™ Ltd., covering several problems dealing with the fine-art logistics. The proposed architecture is an centralized cloud platform processing data collected from a distributed network of wireless sensors capable of edge computing that use extensively IoT and IIoT technologies.

Future developments of the system will allow for full integration for real time premium negotiation based on live risk levels.

REFERENCES

- [1] "Research agenda for the application of ict to cultural heritage," Tech. Rep.
- [2] C. Guccio, M. F. Martorana, I. Mazza, and I. Rizzo, *Technology and Public Access to Cultural Heritage: The Italian Experience on ICT for Public Historical Archives*. Cham: Springer International Publishing, 2016, pp. 55–75. [Online]. Available: https://doi.org/10.1007/978-3-319-29544-2_4
- [3] K. Yang, D. Forte, and M. M. Tehranipoor, "Protecting endpoint devices in iot supply chain," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, ser. ICCAD '15. Piscataway, NJ, USA: IEEE Press, 2015, pp. 351–356. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2840819.2840869>
- [4] M. P. Loria, M. Toja, V. Carchiolo, and M. Malgeri, "An efficient real-time architecture for collecting iot data," in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sept 2017, pp. 1157–1166.
- [5] V. Carchiolo, L. Compagno, M. Malgeri, N. Trapani, M. L. Previti, M. P. Loria, and M. Toja, "An efficient real-time monitoring to manage home-based oxygen therapy," in *Trends and Advances in Information Systems and Technologies*, A. Rocha, H. Adeli, L. P. Reis, and S. Costanzo, Eds. Cham: Springer International Publishing, 2018, pp. 741–749.
- [6] A. Chianese, F. Piccialli, and G. Riccio, "The trust project: Improving the fruition of historical centres through smart objects," *Procedia Computer Science*, vol. 63, pp. 159 – 164, 2015, the 6th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2015)/ The 5th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2015)/ Affiliated Workshops. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1877050915024552>
- [7] Z. He, B. Cui, W. Zhou, and S. Yokoi, "A proposal of interaction system between visitor and collection in museum hall by ibeacon," in *2015 10th International Conference on Computer Science Education (ICCSE)*, July 2015, pp. 427–430.
- [8] M. Buzzi and C. Senette, *RFID Sensors and Artifact Tracking*. Cham: Springer International Publishing, 2017, pp. 435–451. [Online]. Available: https://doi.org/10.1007/978-3-319-50518-3_21
- [9] S. Alletto, R. Cucchiara, G. D. Fiore, L. Mainetti, V. Mighali, L. Patrono, and G. Serra, "An indoor location-aware system for an iot-based smart museum," *IEEE Internet of Things Journal*, vol. 3, no. 2, pp. 244–253, April 2016.
- [10] S. Cuomo, P. D. Michele, F. Piccialli, and A. K. Sangaiah, "Reproducing dynamics related to an internet of things framework: A numerical and statistical approach," *Journal of Parallel and Distributed Computing*, 2017. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0743731517302095>
- [11] R. Mousheimish, Y. Taher, K. Zeitouni, and M. Dubus, "Smart preserving of cultural heritage with PACT-ART - enrichment, data mining, and complex event processing in the internet of cultural things," *Multimedia Tools Appl.*, vol. 76, no. 24, pp. 26 077–26 101, 2017. [Online]. Available: <https://doi.org/10.1007/s11042-017-4900-x>
- [12] R. Mousheimish, Y. Taher, and K. Zeitouni, "Toward the support of challenging service level agreements (slas) in manual and context-dependent activities," in *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2, June 2016, pp. 38–43.
- [13] R. Mousheimish, Y. Taher, K. Zeitouni, and M. Dubus, "Pact-art: Enrichment, data mining, and complex event processing in the internet of cultural things," in *2016 12th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, Nov 2016, pp. 476–483.
- [14] A. K. Hafsa Tahir and M. Junaid, "Internet of things (iot): An overview of applications and security issues regarding implementation," *INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING*, vol. 7, no. 1, 2016.
- [15] G. S. Matharu, P. Upadhyay, and L. Chaudhary, "The internet of things: Challenges amp; security issues," in *2014 International Conference on Emerging Technologies (ICET)*, Dec 2014, pp. 54–59.
- [16] "Furto gioielli: video riprende colpo, in tutto 20 secondi," 2018 (accessed 10/5/2018), http://www.ansa.it/veneto/notizie/2018/01/03/furto-gioielli-a-palazzo-ducale-venezia_41f2d702-6270-4ba3-bae8-47d2b8a03dd2.html.

Information Technology for Management, Business & Society

IT4MBS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the disciplines of information technology and information systems. The IT4BMS area emphasizes the issues relevant to information technology and necessary for practical, everyday needs of business, other organizations and society at large. This area takes a sociotechnical view on information systems and relates also to ethical, social and political issues raised by information systems. Events that constitute IT4BMS are:

- AITM'18—16th Conference on Advanced Information Technologies for Management
- ISM'18—13th Conference on Information Systems Management
- KAM'18—24th Conference on Knowledge Acquisition and Management

AREA SUPERVISORY COMMITTEE

- Carnero Moya, Maria del Carmen, AITSD'18
- Chmielarz, Witold, ISM'18
- Gontar, Beata, IT4L'18
- Komenda, Martin, TEMHE'18
- Korczak, Jerzy, AITM'18
- Pondel, Maciej, KAM'18

16th Conference on Advanced Information Technologies for Management

WE are pleased to invite you to participate in the 16th edition of Conference on “Advanced Information Technologies for Management AITM’18”. The main purpose of the conference is to provide a forum for researchers and practitioners to present and discuss the current issues of IT in business applications. There will be also the opportunity to demonstrate by the software houses and firms their solutions as well as achievements in management information systems.

TOPICS

- Concepts and methods of business informatics
- Business Process Management and Management Systems (BPM and BPMS)
- Management Information Systems (MIS)
- Enterprise information systems (ERP, CRM, SCM, etc.)
- Business Intelligence methods and tools
- Strategies and methodologies of IT implementation
- IT projects & IT projects management
- IT governance, efficiency and effectiveness
- Decision Support Systems and data mining
- Intelligence and mobile IT
- Cloud computing, SOA, Web services
- Agent-based systems
- Business-oriented ontologies, topic maps
- Knowledge-based and intelligent systems in management

EVENT CHAIRS

- **Andres, Frederic**, National Institute of Informatics, Tokyo, Japan
- **Dudycz, Helena**, Wrocław University of Economics, Poland
- **Dyczkowski, Mirosław**, Wrocław University of Economics, Poland
- **Hunka, Frantisek**, University of Ostrava, Czech Republic
- **Korczak, Jerzy**, International University of Logistics and Transport, Wrocław, Poland

PROGRAM COMMITTEE

- **Abramowicz, Witold**, Poznan University of Economics, Poland
- **Ahlemann, Frederik**, University of Duisburg-Essen, Germany
- **Atemezing, Ghislain**, Mondeca, Paris, France
- **Cortesi, Agostino**, Università Ca’ Foscari, Venezia, Italy
- **Czarnacka-Chrobot, Beata**, Warsaw School of Economics, Poland
- **De, Suparna**, University of Surrey, Guildford, United Kingdom

- **Dufourd, Jean-François**, University of Strasbourg, France
- **Franczyk, Bogdan**, University of Leipzig, Germany
- **Januszewski, Arkadiusz**, University of Science and Technology in Bydgoszcz, Poland
- **Kannan, Rajkumar**, Bishop Heber College (Autonomous), Tiruchirappalli, India
- **Kersten, Grzegorz**, Concordia University, Montreal, Canada
- **Kowalczyk, Ryszard**, Swinburne University of Technology, Melbourne, Australia
- **Kozak, Karol**, TUD, Germany
- **Krótkiewicz, Marek**, Wrocław University of Science and Technology, Poland
- **Leyh, Christian**, University of Technology, Dresden, Germany
- **Ligeza, Antoni**, AGH University of Science and Technology, Poland
- **Ludwig, André**, Kühne Logistics University, Germany
- **Magoni, Damien**, University of Bordeaux – LaBRI, France
- **Michalak, Krzysztof**, Wrocław University of Economics, Poland
- **Owoc, Mieczysław**, Wrocław University of Economics, Poland
- **Pankowska, Malgorzata**, University of Economics in Katowice, Poland
- **Pinto dos Santos, Jose Miguel**, AESE Business School Lisboa, Portugal
- **Proietti, Maurizio**, IASI-CNR (the Institute for Systems Analysis and Computer Science), Italy
- **Rot, Artur**, Wrocław University of Economics, Poland
- **Stanek, Stanisław**, General Tadeusz Kosciuszko Military Academy of Land Forces in Wrocław, Poland
- **Surma, Jerzy**, Warsaw School of Economics, Poland and University of Massachusetts Lowell, United States
- **Tazi, El Bachir**, Moulay Ismail University, Meknes, Morocco
- **Teufel, Stephanie**, University of Fribourg, Switzerland
- **Tsang, Edward**, University of Essex, United Kingdom
- **Wątróbski, Jarosław**, University of Szczecin, Poland
- **Wendler, Tilo**, Hochschule für Technik und Wirtschaft Berlin
- **Wolski, Waldemar**, University of Szczecin, Poland
- **Zanni-Merk, Cecilia**, INSA de Rouen, France
- **Ziamba, Ewa**, University of Economics in Katowice, Poland

Business Process Management: Terms, Trends and Models

Renato Neder, Paulo Ramalho, Olivian Rabelo, Elisandra Zambra
Faculty of Management, Federal University of Mato Grosso, Cuiabá, MT, Brazil
Email: renatoneder, paramalho, olivanrabelo, elisandrazambra@gmail.com

Cristiano Maciel
Institute of Computing, Federal University of Mato Grosso, Cuiabá, MT, Brazil
Email: crismac@gmail.com

Nathalia Benevides
Faculty of Management, Federal University of Mato Grosso, Cuiabá, MT, Brazil
natyfbenevides@gmail.com

Abstract—Understanding the Business Process Management (BPM) subject is a complex and multifaceted task, which is why the goal of this study is to explore the scientific production concerning BPM in its many dimensions. In order to do so, the proposed methodology is quantitative, bibliometric and longitudinal, and uses the Semantic Network Analyses as a way to explore a large set of scientific documents. The foundation for this research included 765 articles about BPM. This study managed to limit the subject of BPM to the fields of Business Administration and Information Technology and estimate research trends in both fields. The quantitative method employed in this research study is considered to be a limiting factor, because it does not permit large volumes of data to be analyzed, albeit with little depth. The achieved results allow IT and Business Administration to understand the dynamics of the scientific production network about BPM, in addition to identifying research trends in its field of study. Analysis of the BPM dimensions is innovatively achieved from the standpoint of semantic networks.

I. INTRODUCTION

BUSINESS Process Management (BPM) has the potential to support organizational changes, since it shifts the focus from managing functional areas (departments) to business processes. This paradigmatic shift can allow managers to organize efforts around tasks, flows and people in order to improve client delivery. BPM, under a technical and pragmatic perspective, can serve as a tool to adjust the organization towards its managerial strategy.

BPM most likely has its roots in Total Quality Management (TQM), a Japanese quality model developed in post-war 1940s. According to Capote [1], the most evident principle of TQM was the need to establish a shared organizational consciousness about the importance of high quality managerial and productive processes involving external elements to the original control mechanisms of the organization, such as the relationship between suppliers and the remaining parties involved in developing the business. Thus, TQM incorporated the need to understand processes as a set of tasks, implemented by machines or people, with the potential of improving continuously.

Due to its characteristics, BPM provides organizational control at the level of processes, tasks, activities and individ-

uals. In this sense, organizations that implement it seek control of their processes, which makes process modeling an important predictor of government policies and accountability. BPM represents the shift from a vertical and hierarchical hegemonic paradigm to a horizontal paradigm that integrates multiple business functions. This fact has justified the increased interest of researchers and consultants regarding the topic, which can be noted from the growing number of scientific articles and academic studies regarding BPM.

In order to have an idea of the increase in publications about the topic, when one types “business process management” in Google Scholar, approximately 139,000 results are found. When searching the same term in the Web of Science database, approximately 2,435 documents are found. If the growth of scientific publications in the Web of Science database is analyzed, one verifies that only two articles existed in 1994; 45 in 2004; 208 in 2014 and 283 in 2016; which confirms the increase in scholarly production in the field [2].

The plurality of viewpoints, in addition to the interdisciplinary nature of BPM discussions, broadens the complexity involved in the subject, thus rendering it harder to comprehend. When searching for the term “business process management” in the Web of Science database, many “categories” are displayed, such as computer science information systems; computer science artificial intelligence; computer science theory method; computer science interdisciplinary applications; computer science software engineering; management; business.

Besides the categories employed, others were also present, such as: engineering; economics; operations research, among others, which corroborates with the interdisciplinary nature of the topic in various study areas. For the purpose of this research, categories presented by the Web of Science database were simplified and, for this reason, IT and Management categories were restricted to two fields of knowledge (Management and Information Technology).

Due to the increase in the volume of worldwide scholarly production regarding BPM, we observed that the traditional research methods that involve reading, indexation and manual analysis of scientific documents have not been

sufficient to deal with the subject's growing complexity. On the other hand, advances in computer technology, especially with algorithms that allow data mining and semantic text analysis, may support researchers in the task of dealing with extensive worldwide scientific production, thus optimizing research resources.

Therefore, the objective of this study is to analyze scientific production in BPM in its various dimensions and, more specifically, understand the existing relations between the fields of Information Technology and Management, moreover to establish models and estimate trends in these fields. From a methodological viewpoint, this research has a bibliographical quality, with a quantitative, bibliometric, and longitudinal approach. The data used are from secondary sources from studies extracted in the Web of Science database and semantic relationships are analyzed among the articles selected by means of the Semantic Network analysis method.

II.2 THEORETICAL REFERENCE

The foundation for this research requires a discussion of BPM in two fields related to the present study: Management and Information Technology.

A. BPM in the Field of Management

Business Process Management, understood as a management theme, comprehends several dimensions, such as: organization culture; organization performance; organization conduct; corporate governance; and competitive advantage. Thus, it is crucial to present basic concepts concerning these dimensions.

BPM and organization culture: Culture can be defined as a set of shared values within a group, manifested through ideas, attitudes, rituals, technologies, products and institutions. These values can vary from group to group or from institution to institution and are defined as ideas that influence the group's behavior and organize the group's model [3]. For Schein [4], group culture may be defined as a pattern of assumptions that have been instituted by a given group, in the sense that it solves the problems of adaptation and internal integration, which worked well enough to be considered valid and, therefore, to be taught to new members as the correct way to perceive, think about and feel regarding these problems. In regard to business management, Vukšić et al. [5] state that organizational culture can be understood as an organizational style that reveals the personality of the organization and determines the actions and behavior of collaborators. The study by vom Brocke and Sinnl [6] presents three associations between the terms "culture" and "BPM", as follows: a) culture as an independent factor that influences BPM; b) culture as a dependent factor influenced by BPM; and c) culture as equivalent to BPM culture. According to Santos [7], BPM practices involve a deep analysis of the organization and changes in its organizational structure. Some organizations have a culture that may be incompatible with the desire to organize itself around the

client and, consequently, of business process management. Therefore, methodological support should be conducted to continuously improve its business processes, coordinate activities, define precisely the responsibilities of each person involved, and create a process office. Changing the culture of a function-centered organization to a process-centered culture is a big challenge for BPM initiatives to gain space in the organizational context and to produce the expected results and performance [8]. For Baumöl [9], one of the critical factors of success in company transformation by means of process changes is the receptivity of people and its commitment to new forms of doing things.

BPM as support to organization conduct: A line of studies on BPM highlights the impact of process management in organizational strategies. In this context, the result of organizational conduct tends to present strategic gains through the rationalization of organizational processes. However, research studies highlight the complexity involved in developing mechanisms that provide an alignment between BPM strategies and the result of organizational conduct, given the complexity associated with multi-criteria evaluation structures of stakeholder indicators. Chang [10] highlights that BPM was originally a process-oriented organizational approach used to project, analyze and perfect business processes for managing and improving organizational conduct more efficiently. The positive relationship between BPM practices and the result of organizational conduct is a recurrent affirmation in a set of research studies in different strategic business models [11-12-13-14-15]. The relationship between BPM and the result of organizational conduct is so deep that authors Dijkman et al. [13] emphasize that BPM maturity tends to improve sustainability of organizational conduct in the long run.

BPM and Organization Performance: The performance evaluation methodologies based on organizational processes could give support to BPM propagation within organizations, creating visible commercial processes by measuring intermediary and final results. The benefits of having a measuring system will decrease with time if they do not reflect the changes in organizational processes. If there is a lack of focus and update in business processes, the decision-makers and the main interested parties will find it hard to determine if the company is achieving continuous progress towards its strategic goals [14]. A measuring system is crucial in management processes for providing assessment and spreading success stories for motivational purposes, evaluating progress, allocating and redistributing resources, and instilling a continuous improvement system for the ES life cycle [17]. The performance measured in an organization can change rapidly (a typical example would be a drastic change in a bank's results before/during a financial crisis) and, therefore, the performance at the organizational level is not a good proxy for measuring BPM success. Besides, the success of a BPM is hard to define depending on the specific goals of each project and no single document has provided a wide-ranging definition [18]. Finally, Bititci et al. [19] state that any performance assessment system should be balanced and integrated.

BPM as a tool for supporting corporate governance: Contemporary organizations are encouraged to adopt strategies that implement positive differentiation in segments in which they work if they wish to maintain itself and survive in the market. Factors such as elevated competitiveness between companies, as well as fast and constant changes in the type of client that is increasingly close to organizational limits, render the business environment even more complex. This scenario leads businesses to adopt new management techniques, the most relevant one being Business Process Management (BPM). According to Jesus et al. [20], the main advantages of BPM are: process autonomy, improvement of performance monitoring, redefinition of organizational structure, and implementation of reference models. Governance is inserted in the BPM context as an efficient approach in its implementation, aligned with process management branches of organizations, because it helps to migrate isolated BPM initiatives to integrated and synergic initiatives. This integrative idea of business process modelling demonstrates the importance of creating structures of process implementations based on governance, with the perspective of providing greater involvement of the participants and transparency throughout all the processes. The concept of process governance is associated with the creation of relevance and transparency in relation to responsibility, decision making and reward system in order to guide actions [21]. BPM governance is articulated with goals, principles, and organograms that reveal who is allowed to make decisions, as well as policies and norms that define what managers will do [22]. There is a need to adapt governance to what BPM demands. This is the foundation for adding value by means of agility and scalability.

BPM as a tool for competitive advantage: The main objective of BPM, in this perspective, is to create competitive advantage to the company, thus guaranteeing quality of products and services, satisfying the customer with delivery that is superior to the competition. In this sense, it acts as a tool for competitive advantage, in a continuous effort towards process improvement [16]. In order for the company to achieve competitive advantage over its competitors, Barney [23], Dierickx and Cool [24] and Dyer [25] state that they will have to be able to accumulate resources and abilities that are valuable, non-replaceable and hard to emulate. Concerning sustainable competitive advantage, Brito and Vasconcelos [26] state that organizational resources should be rare (not easily available to other companies), difficult to emulate, and the company should possess organizational conditions to explore resources. In this sense, Molardi and Pontes [27] show that business process management may generate competitive advantage for the organization, since they directly affect management and add value to the product or service delivered to the client. In that light, BPM can dynamize and guarantee efficiency of organizational resources, thus proving to be an important tool in seeking sustainable competitive advantage.

B. BPM in the field of Information Technology

When analyzing business process management as an IT discipline, other dimensions and categories appear as a differential when compared to BPM discussions in the field of business management. In this sense, it is necessary to present differences.

IT management has become an increasingly important issue in organizations and in the academic-scientific realm. In this sense, process management emerges as a management strategy that meets the needs of IT, as it originates from an effort to manage organizations through their business processes. In this sense, several perspectives arise when associating IT with BPM. The main function of information technology, in this perspective, is to enable the performance of an organization's business processes in order to create value for customers and shareholders. It is also observed that virtually all process improvement initiatives rely on IT support [28]. For Rahimi et al. [29] the association between business process management and IT management, based on the analysis of academic literature about business processes and IT capabilities, is found to be under-explored. BPM can thus collaborate to fill an existing gap between business fields and Information Technology fields.

A key concept to address this gap is governance. According to Haes and Van Grembergen [30] IT governance is a priority in the agenda and several organizations are implementing IT governance practices in day-to-day operations. The authors suggest that organizations are beginning to implement IT governance in order to achieve better alignment between business and IT.

According to Spanyol [28], in order to optimize and maintain organizational performance improvements, some form of governance is needed to create appropriate structures, measures, roles, and responsibilities to assess and manage end-to-end business process performance. The author adds that one of the roles of governance is to ensure that IT investments are closely associated with the organization's business strategy and that IT investment offsets come from specific improvements in business process performance. Other authors, such as Reijers [31] and Ramesh et al. [32] study BPM from the perspective of business process management system (BPMS) implementation. For them, the success of BPM deployment strategies is closely related to the technologies involved in this process and with the aim of predicting the success of BPMS implementation based on the maturity of the level of understanding of processes within an organization.

Dumas and Kohlborn [33], in turn, bring the concept of service-oriented architecture (SOA) in the context of BPM, presenting a method to analyze a process so it can be executed in the context of an application. The authors bring SOA as a computational paradigm in order to use distributed capabilities. It is important to highlight that the sense of capabilities, in this context, refers to both the capabilities offered by the business and those offered by specific application systems.

III METHODOLOGICAL PROCEDURES

The methodology proposed for this study is bibliographical, with a quantitative, bibliometric and longitudinal approach. The data used are from secondary research sources extracted from the Web of Science database. The semantic relationships among 661 articles were analyzed through the Semantic Network analysis method.

According to Lopes [34] a database search strategy can be defined as a set of rules and techniques that make it possible to find the desired information stored in a database. The author points out that, in order to achieve the desired response by the researcher, it is necessary to perform logical operations, by restricting the results achieved or by expanding them to obtain information that may be relevant to the research.

The following research restriction was considered for this article: (Topic: ("business process management")); refined by types of document: (article) and categories from Web of Science: (computer science information systems or computer science artificial intelligence or management or computer science theory methods or computer science software engineering or business or operations research management science or computer science interdisciplinary applications); and estimated period: all years with indexes: sci-expanded, ssci, a&hci, cpci-s, cpci-ssh, esci.

A Text mining

In the data mining and cleaning phase, the CASOS institute's AutoMap software was used. The research database was divided in two; the former presented articles that shared more relevance to Business Management, and the latter contained articles that were more associated with Information Technology. Both databases were subsequently submitted to text mining processes using the Automap software.

As for the summary of the mining steps, the Perform All Cleaning and Perform All Preparation algorithms were first executed, then numbers, pronouns, prepositions, punctuation and symbols were deleted and the text was converted to upper case. Then, a list of concepts was generated that created another list of eliminations, in order to exclude concepts that did not appear in at least three articles. The next step was to create a list of bigrams using the TF-IDF metric. One hundred bigrams with the highest numbers were selected for this metric. After the bigrams were chosen, all other concepts were excluded from the analysis and the networks were generated. Finally, the networks were analyzed in the ORA software and the following reports were analyzed: Network Comparison and All Measures by Category

B. Analysis of semantic networks

An aspect of Social Network Analysis that has been highlighted in the academic community in recent years is the Semantic Network Analysis. According to Atteveltdt [35], it is possible to define Semantic Network Analyses as the

analysis of a thematic content in which the messages are deconstructed into semantic units that, in turn, are diluted into one or more variables, which are then recomposed through combination and aggregation techniques.

Semantic networks, according to Sowa [36], are considered to be structures of knowledge representation that are formed by vertices and edges. One can understand a semantic network as one in which its nodes present semantic content, or "meaning". Lee et al. [37] considers the semantic network as a concept graph. The Semantic Network Analysis (SeNA) can be considered an extension of social network analysis (SNA) that explores relationships among meanings shared in linguistic and social configurations. In order to understand the importance of a unit of meaning in a semantic network, metrics are employed.

According to Gloor et al. [38], the SeNA conducts a time-based calculation of network centrality measures, social network visualizations, as well as semantic process of text mining, cleaning and analysis. Back to Lee et al. [37], a semantic network analysis is a part of network analysis that explores the relationships between meanings shared in linguistic settings. The analysis of semantic networks can, like social network analysis, be executed by several metrics, such as: network density, degree of centrality, centrality betweenness, eigenvector centrality, path length, among others.

C. Metrics

This article will use centrality metrics, which roughly identifies the relative importance of nodes in a network. Thus, the greater the centrality metric, the greater the importance of this node in the network. It is possible to define centrality as the property of a node or a group of nodes that relate to its position in a network [39]. For the authors, thinking in terms of centrality means trying to understand the contribution that a node or a set of nodes offer to the structure of this network; in other words, centrality is the degree of structural importance of a node in relation to the network. In order to estimate the time trends in networks, the Betweenness Centrality Newman [40] metric is a measure of the centrality of a node in a network usually calculated as the fraction of shorter paths between pairs of nodes passing through the node of interest. The metric betweenness centrality, for Chen et al. [41], can be defined individually for each network node, as measured in the degree to which the node is in the middle of the path that connects it to the other vertices of this network.

IV. RESULTS AND DISCUSSIONS

In this section, the collected data are presented and discussed.

A. Comparison between Management and Information Technology networks

By means of the algorithms outputs “Network Comparison” of the ORA software, the relationships are identified between the two networks which concern this study.

Table 01 and 02 summarize pieces of information about the two networks and their relationships. It should also be noted that the Management Network has a total of 148 texts and 95 concepts, while the Information Technology presents 513 texts and 97 concepts.

The Management and IT network density levels shows that the Management network density is 0.066. For Valente [42], low-density networks ($R > 0.100$) can have limited efficacy regarding the concept flow. Table 1 presents the proximity of semantic content consolidated in analyses.

The measure demonstrates the closeness of semantic content (represented by network nodes) and its links. All identified values present numbers higher than 90%.

The Common Focus report (Software ORA) presented the following bigrams: bpm-capability, business-environment, business-process, business-system, competitive-advantage decision-make management-bpm, management-system, maturity-model performance-measurement, process-description, process-design, process-execution, process-improvement, process-knowledge, process-management, process-mine, process-model, process-monitor service oriented-architecture, social-software web-service, workflow-process

B. Management Network

The analysis of the network allows a glance at the BPM / Management network concepts and the existing links between them, as well as the elements that are external to the network that are the concepts of less centrality. The network is composed of 148 article abstracts, has 95 nodes, a density of 0.043 and 588 links.

Table 1 presents the evolution of the 16 bigrams with the largest Betweenness Centrality in the network from 1995 to 2018 compared to previous periods. Note that positive, negative, neutral trends and new perspectives have been identified.

Table 1. Evolution of Betweenness Centrality for BPM terms - business management.

Concepts/Period/ Centrality- Betweenness	1st Period: 2006- 1995	2nd Period: 2010- 1995	3rd Period: 2014- 1995	4th Period: 2018 - 1995	Trends
business-process	0.92	0.839	0.611	0.354	Negative
management-bpm	0.009	0.108	0.213	0.243	Positive
process- management	0.054	0.139	0.051	0.042	Neutral
process- performance	x	x	0.044	0.041	Positive
decision-make	0.114	0.02	0.034	0.04	Negative
organizational- performance	0	0.02	0.081	0.034	Positive
supply-chain	0.059	0.04	0.03	0.033	Neutral
management- system	0.114	0.078	0.013	0.028	Negative
knowledge- management	x	0	0.043	0.026	Positive
process-model	0.059	0.041	0.038	0.025	Negative
business- environment	0.06	0.055	0.049	0.023	Negative
operation- management	0	0	0.003	0.023	Positive
managerial-system	x	x	0.056	0.023	Negative
traditional-bpm	x	x	x	0.023	New perspective
process- improvement	x	0.019	0.013	0.018	Neutral
maturity-model	x	x	7.23E- 04	0.016	Neutral

Source: Research Data

The term "New perspective" describe bigrams that only appeared in the network in the last period: conceptual-model and traditional-bpm. Positive trends in the Management network identified the following concepts: management-bpm; process-performance; organizational-performance; knowledge-management; operation-management. Neutral trends are represented by the following concepts: process-management; supply chain; process-improvement; maturity-model; bpm-initiative.

C. Information Technology Network

The construction of the IT network was based on a set of 513 article abstracts from the database found in Web of Science. It presented 97 nodes, a density of 0.037 and 950 links.

Table 2 presents the 16 main network concepts according to the metric Betweenness Centrality and its development over time. As positive trends in the Information Technology network, the methodology used presented the concepts: business-process, erp-system, decision-make, management-system, business-environment, process-model, serviceoriented-architecture and web-service. The new perspectives presented were: quality-management, process-

logic, continuous-improvement, process-orientation, supply-chain, and process-quality.

V.FINAL CONSIDERATIONS

In a global context in which there is a massive production of scientific papers, traditional methods of analysis, which require reading and cognitive interpretation of scientific production, are not sufficient for analyzing large volumes of documents with limited research resources. As an alternative to traditional methods, new ways of analyzing complex world production arise with the aid of text mining and semantic content analysis software. In this sense, this study offers to analyze the scientific production of BPM articles in the fields of Management and Information Technology, in their specificities and in their commonalities. Regarding the comparative analysis between the Management and Information Technology networks, no significant differences were verified, as seen in Tables 1. As a common focus between the two networks, the analyses have led to the following concepts: bpm-capability, business-environment, business-process, business-system, competitive-advantage decision-make management-bpm, management-system, maturity-model performance-measurement, process-description, process-design, process-execution, process-improvement, process-knowledge, process-management, process-mine, process-model, process-monitor service-oriented-architecture, social-software web-service, workflow-process.

Regarding the networks' singularities, the analyses demonstrate key concepts for being understood. In the Management network, one can observe these concepts through Figure 1, which presents the result of an algorithm that brings the three metrics with the highest value per concept, and Table 2, which presents a ranking of the 16 largest values for the Betweenness Centrality.

In order to meet the goal of tracing research trends in semantic networks, the 16 concepts with the greatest centrality in the period from 2008 to 2015 were analyzed in relation to the Betweenness Centrality metric in relation to the previous periods, according to Table 2. This analysis brought the following concepts with positive trends: management-bpm; process-performance; organizational-performance; knowledge-management; operation-management.

In the Information Technology network, Figure 2 presents the most important concepts according to an algorithm of the ORA software. Furthermore, Table 3 presents the 16 concepts with the highest Betweenness Centrality metric to the IT network. As for the positive trends in the IT network, there are the following concepts: business-process, erp-system, decision-make, management-system, business-environment, process-model, service-oriented-architecture and web-service.

The IT network also presented six new perspectives represented by the following concepts: quality-management [46][47], process-logic [48][49][50], continuous-improvement

[51][52][53], process-orientation [54][55], supply-chain [15][56][57], and process-quality [58][59]. For management, only the traditional-bpm concept was represented [43][44][45].

As a methodological contribution, this study structures an efficient way to analyze a large number of scientific documents in any area of knowledge. However, considering the research limitations, it is observed that the method used to analyze a large number of scientific articles does so with little depth and thus should be used in association with traditional research methods. In this sense, a suggestion for future research is to conduct a traditional bibliographical study of the new perspectives and positive trends found in the analyses. Furthermore, conducting this study with other foundations is also recommended.

Finally, it is important to emphasize the relevance of studies such as this one, which synthesizes a large set of data about the field, which can be used by researchers and industry for reflections on BPM in the fields of Management and IT, leading to the emergence of trends in these fields.

ACKNOWLEDGMENT

We would like to thank the Federal University of Mato Grosso (UFMT), Mato Grosso State Audit Court (TCE-MT) and Uniselva Foundation for all the support to develop and publish this research.

REFERENCES

- [1]. Capote G (2012) BPM para todos: uma visão geral abrangente, objetiva e esclarecedora sobre gerenciamento de processos de negócio [BPM for Everyone: A Comprehensive, Objective, and Enlightening Overview of Business Process Management]. Gart Capote, Rio de Janeiro. Portuguese.
- [2]. Web of Science (2018) Web of Science Core Collection Home. <https://webofknowledge.com>
- [3]. Schmiedel T, vom Brocke J, Recker J (2014) Development and validation of an instrument to measure organizational cultures support of Business Process Management. *Information & Management* 51(1):43-56. doi:10.1016/j.im.2013.08.005
- [4]. Schein EH (2004) *Organizational culture and leadership* (3rd ed.) John Wiley & Sons, San Francisco
- [5]. Vukšić VB, Vugec DS, Lovrić A (2017) Social business process management: croatian it company case study. *Business Systems Research Journal* 8(1):60-70. doi:10.1515/bsrj-2017-0006
- [6]. vom Brocke J, Sinnl T (2011) Culture in business process management: a literature review. *Business Process Management Journal* 17(2):357-378. doi:10.1108/1463715111122383
- [7]. Santos HRM (2012) Fatores críticos de sucesso das iniciativas de BPM no setor público [Critical success factors of BPM initiatives in the public sector] [master's thesis]. Recife (PE): Universidade Federal de Pernambuco. Portuguese. <https://repositorio.ufpe.br/handle/123456789/10877>
- [8]. Soso FA (2016) Fatores que caracterizam a adoção do Business Process Management (BPM) pelas organizações [Factors that characterize the adoption of Business Process Management (BPM) by organizations] [master's thesis]. São Leopoldo (RS): Universidade do Vale do Rio Sinos. Portuguese. <http://www.repositorio.jesuita.org.br/handle/UNISINOS/5592>
- [9]. Baumöl U (2013) Mudança cultural na gestão de processos. In: vom Brocke J, Roseman M (eds) *Manual de BPM: gestão de processos de negócio* [BPM Handbook: Business Process

- Management]. Bookman Editora, Porto Alegre, pp 331-358. Portuguese.
- [10]. Chang JF (2006) Business process management systems: strategy and implementation. Auerbach Publications, New York
- [11]. Kumar V, Smart PA, Maddern H, Maull RS (2008) Alternative perspectives on service quality and customer satisfaction: the role of BPM. *International Journal of Service Industry Management* 19(2):176-187. doi:10.1108/09564230810869720
- [12]. Smart PA, Maddern H, Maull RS (2009) Understanding business process management: implications for theory and practice. *British Journal of Management* 20(4):491-507. doi:10.1111/j.1467-8551.2008.00594.x
- [13]. Dijkman R, Lammers SV, Jong A (2015) Properties that influence business process management maturity and its effect on organizational performance. *Information Systems Frontiers* 18(4):717-734. doi:10.1007/s10796-015-9554-5
- [14]. Pádua SID, Jabbour CJC (2015) Promotion and evolution of sustainability performance measurement systems from a perspective of business process management. *Business Process Management Journal* 21(2):403-418. doi:10.1108/bpmj-10-2013-0139
- [15]. Pradabwong J, Braziotis C, Tannock JDT, Pawar KS (2017) Business process management and supply chain collaboration: effects on performance and competitiveness. *Supply Chain Management: An International Journal* 22(2):107-121. doi:10.1108/scm-01-2017-0008
- [16]. Trkman P (2010) The critical success factors of business process management. *International Journal of Information Management* 30(2):125-134. doi:10.1016/j.ijinfomgt.2009.07.003
- [17]. Al-Mudimigh AS (2007) The role and impact of business process management in enterprise systems implementation. *Business Process Management Journal* 13(6):866-874. doi:10.1108/14637150710834604
- [18]. Škriňjar R, Trkman P (2013) Increasing process orientation with business process management: Critical practices. *International Journal of Information Management* 33(1):48-60. doi:10.1016/j.ijinfomgt.2012.05.011
- [19]. Bititci U, Cavalieri S, Cieminski G (2005) Implementation of performance measurement systems: private and public sectors. *Production Planning and Control* 16(2):99-100. doi:10.1080/09537280512331333002
- [20]. Jesus L, Macieira A, Karrer D, Caulliraux H (2013) Escritório de processos: estudo de caso sobre uma empresa Brasileira. In: vom Brocke J, Roseman M (eds) *Manual de BPM: gestão de processos de negócio [BPM Handbook: Business Process Management]*. Bookman Editora, Porto Alegre, pp 307-328. Portuguese.
- [21]. Richardson C (2006) Process governance best practices: building a BPM center of excellence. *Business Process Trends*, 1-6. <https://www.bptrends.com/bpt/wp-content/publicationfiles/09-06-ART-ProcessGovernanceBestPractices-Richardson1.pdf>
- [22]. Harmon P (2007) Business process change: a guide for business managers and BPM and Six sigma professionals. Morgan Kaufmann, Boston. doi:10.1016/b978-012374152-3/50043-4
- [23]. Barney J (1991) Firm resources and sustained competitive advantage. *Journal of Management* 17(1):99-120. doi:10.1177/014920639101700108
- [24]. Dierickx I, Cool K (1989) Asset stock accumulation and sustainability of advantage. *Management Science* 35(12):1514-1514. doi:10.1287/mnsc.35.12.1514
- [25]. Dyer, JH (1996) Specialized supplier networks as a source of competitive advantage: evidence from the auto industry. *Strategic Entrepreneurship Journal* 17(4):271-291. doi: 10.1002/(sici)1097-0266(199604)
- [26]. Brito LAL, Vasconcelos FC (2004) A Heterogeneidade do desempenho, suas causas e o conceito de vantagem competitiva: proposta de uma métrica [Heterogeneity of performance, its causes and the concept of competitive advantage: proposal of a metric]. *Revista de Administração Contemporânea* 8(spe):107-129. Portuguese. doi:10.1590/s1415-65552004000500007
- [27]. Molardi RM, Pontes AT (2017) Fatores críticos de sucesso em iniciativas de BPM na administração pública [Critical Success Factors in BPM Initiatives in Public Administration]. In: *Proceedings of the 20th Seminário em Administração da Universidade de São Paulo-FEA/USP. SemeAd, São Paulo*, pp 1-17. Portuguese. <http://login.semead.com.br/20semead/arquivos/1460.pdf>
- [28]. Spanyi A (2013) Governança de BPL. In: vom Brocke J, Roseman M (eds) *Manual de BPM: gestão de processos de negócio [BPM Handbook: Business Process Management]*. Bookman Editora, Porto Alegre, pp 261-277. Portuguese.
- [29]. Rahimi F, Møller C, Hvam L (2016) Business process management and IT management: the missing integration. *International Journal of Information Management* 36(1):142-54. doi:10.1016/j.ijinfomgt.2015.10.004
- [30]. De Haes S, Van Grembergen, W (2009) An exploratory study into IT governance implementations and its impact on business/IT Alignment. *Information Systems Management* 26(2):123-137. doi:10.1080/10580530902794786
- [31]. Reijers HA (2006) Implementing BPM systems: the role of process orientation. *Business Process Management Journal* 12(4):389-409. doi:10.1108/14637150610678041
- [32]. Ramesh B, Jain R, Nissen M, Xu P (2005) Managing context in business process management systems. *Requirements Engineering* 10(3):223-237. doi:10.1007/s00766-005-0005-6
- [33]. Dumas M, Koheborn T (2013) Gestão de processos habilidades por serviços. In: vom Brocke J, Roseman M (eds) *Manual de BPM: gestão de processos de negócio [BPM Handbook: Business Process Management]*. Bookman Editora, Porto Alegre, pp 151-171. Portuguese.
- [34]. Lopes IL (2002) Uso das linguagens controlada e natural em bases de dados: revisão da literatura [Use of controlled and natural languages in databases: literature review]. *Ciência da Informação* 31(1):41-52. Portuguese. doi:10.1590/s0100-19652002000100005
- [35]. Atteveldt W Van (2008) *Semantic Network Analysis; Techniques for Extracting, Representing, and Querying Media Content*. [PhD thesis]. Amsterdam: Vrije Universiteit.
- [36]. Sowa JF (1991) *Principles of semantic networks*. Morgan Kaufmann Publishers, California
- [37]. Lee H, Lee DI, Kim T, Lee J (2013) The moderating role of socio-semantic networks on online buzz diffusion. *Journal of Business Research* 66(9):1367-1374. doi:10.1016/j.jbusres.2012.02.038
- [38]. Gloor PA, Krauss J, Nann S, Fischbach K, Schoder D (2009) Web Science 2.0: Identifying Trends through Semantic Social Network Analysis. In: *IEEE International Conference on Computational Science and Engineering*. CSE, Vancouver, pp 215-222. doi:10.1109/cse.2009.186
- [39]. Borgatti PS, Everett GM, Johnson JC (2013) *Analyzing social networks*. SAGE Publications Limited, London
- [40]. Newman MEJ (2005) A measure of betweenness centrality based on random walks. *Social Networks* 27(1):39-54. doi:10.1016/j.socnet.2004.11.009
- [41]. Chen C, Sanjuan FI, Hou J (2010) The structure and dynamics of cocitation clusters: a multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology* 61(7):1.386-1.409. doi:10.1002/asi.21309
- [42]. Valente TW (1995) *Network models of the diffusion of innovations*. Hampton Press, Cresskill
- [43]. Afflerbach P, Hohendorf M, Manderscheid J (2017) Design it like Darwin-A value-based application of evolutionary algorithms for proper and unambiguous business process redesign. *Information Systems Frontiers* 19(5):1101-1121. doi:10.1007/s10796-016-9715-1
- [44]. Chang C, Srirama SN, Buyya R (2016) Mobile Cloud Business Process Management System for the Internet of Things. *ACM Computing Surveys* 49(4):1-42. doi:10.1145/3012000

- [45] Suša Vugec D, Tomičić-Pupek K, Vukšić VB (2018) Social business process management in practice. *International Journal of Engineering Business Management* 10, 184797901775092. doi:10.1177/1847979017750927
- [46] Jakhar SK (2016) Stakeholder Engagement and environmental practice adoption: the mediating role of process management practices. *Sustainable Development* 25(1):92–110. doi:10.1002/sd.1644
- [47] Malfait S, Van Hecke A, Hellings J, De Bodt G, Eeckloo K (2016) The impact of stakeholder involvement in hospital policy decision-making: a study of the hospital's business processes. *Acta Clinica Belgica* 72(1):63–71. doi:10.1080/17843286.2016.1246681
- [48] Delgado A, Calegari D, Arrigoni A (2016) Towards a Generic BPMS User Portal Definition for the Execution of Business Processes. *Electronic Notes in Theoretical Computer Science* 329: 39–59. doi:10.1016/j.entcs.2016.12.004
- [49] Martin N, Depaire B, Caris A (2015) The Use of Process Mining in Business Process Simulation Model Construction. *Business & Information Systems Engineering* 58(1):73–87. doi:10.1007/s12599-015-0410-4
- [50] Teixeira S, Agrizzi BA, Filho JGP, Rossetto S, Baldam RL (2017) Modeling and automatic code generation for wireless sensor network applications using model-driven or business process approaches: A systematic mapping study. *Journal of Systems and Software* 132:50–71. doi:10.1016/j.jss.2017.06.024
- [51] Bisogno S, Calabrese A, Gastaldi M, Levialdi Ghiron N (2016) Combining modelling and simulation approaches. *Business Process Management Journal* 22(1):56–74. doi:10.1108/bpmj-02-2015-0021
- [52] Johannsen F, Fill HG (2017) Meta Modeling for Business Process Improvement. *Business & Information Systems Engineering* 59(4):251–275. doi:10.1007/s12599-017-0477-1
- [53] Lehnert M, Linhart A, Roeglingle M (2017) Exploring the intersection of business process improvement and BPM capability development. *Business Process Management Journal* 23(2):275–292. doi:10.1108/bpmj-05-2016-0095
- [54] Khosravi A (2016) Business process rearrangement and renaming. *Business Process Management Journal* 22(1):116–139. doi:10.1108/bpmj-02-2015-0012
- [55] Nadarajah D, Syed A, Kadir SL (2016) Measuring Business Process Management using business process orientation and process improvement initiatives. *Business Process Management Journal* 22(6):1069–1078. doi:10.1108/bpmj-01-2014-0001
- [56] Palattella MR, Dohler M, Grieco A, Rizzo G, Torsner J, Engel T, Ladid L (2016) Internet of Things in the 5G Era: Enablers, Architecture, and Business Models. *IEEE Journal on Selected Areas in Communications* 34(3):510–527. doi:10.1109/jsac.2016.2525418
- [57] Yousfi A, Bauer C, Saidi R, Dey AK (2016) uBPMN: A BPMN extension for modeling ubiquitous business processes. *Information and Software Technology* 74:55–68. doi:10.1016/j.infsof.2016.02.002
- [58] Faroque AR, Morrish SC, Ferdous AS (2017) Networking, business process innovativeness and export performance: the case of South Asian low-tech industry. *Journal of Business & Industrial Marketing* 32(6):864–875. doi:10.1108/jbim-06-2015-0113
- [59] Kalenkova AA, Van Der Aalst WMP, Lomazova IA, Rubin VA (2015) Process mining using BPMN: relating event logs and process models. *Software & Systems Modeling* 16(4):1019–1048. doi:10.1007/s10270-015-0502-0

Development of crowd investing on the basis of ICO crypto assets using block-options for the supply of electric generation capacity

Andrew Varnavskiy
Financial University under the
Government of the RF,
Leningradsky Prospekt 49,
Moscow, Russia
Email: AVVarnavskiy@fa.ru

Ul'ia Gruzina
Financial University under the
Government of the RF,
Leningradsky Prospekt 49,
Moscow, Russia
Email: ymgruzina@fa.ru

Artur Rot
Wroclaw University
of Economics,
ul. Komandorska 118/120,
53-345 Wroclaw, Poland
Email: artur.rot@ue.wroc.pl

Vladislav Trubnikov
Financial University under the
Government of the RF,
Leningradsky Prospekt 49,
Moscow, Russia
Email: vladtrubnikov95@gmail.com

Anastasia Buryakova
Financial University under the
Government of the RF,
Leningradsky Prospekt 49,
Moscow, Russia
Email: AOBuryakova@fa.ru

Ekaterina Sebechenko
Financial University under the
Government of the RF,
Leningradsky Prospekt 49,
Moscow, Russia
Email: EVSebechenko@fa.ru

Abstract—attraction of investments into the electric power industry is complicated by a number of problems related to the long payback period and instability of the conditions on the market. Investors in the electric power industry must invest huge sums of money and hope for maintaining high demand and prices in the future in order to get a payback from the project. Decentralized investment distributes risks and allows you to raise a sufficient profit. In the paper Authors will consider the possibility of using distributed register systems to involve potential energy consumers in investing in the construction of generating facilities with a certain amount of energy at a reduced price in the future. In addition, the blockchain system allows to solve problems of the electric power market: simplify and make more flexible maintenance of transactions, automate trade settlements, reduce the risks of non-fulfillment of obligations. In turn, the improved interaction environment of market participants will make the market more stable, which will increase the investment attractiveness of the industry.

I. INTRODUCTION

According to experts, the provision of energy generation facilities by 2023-2027 may be exhausted due to the growing demand for energy, as well as the depreciation of previously constructed facilities [1]. Existing financial instruments do not allow a broad mass of investors to participate in financing the energy sector: these are mostly utilized by professional participants. At the same time, future energy consumers are interested in acquiring it at reduced tariffs and could become anchor investors in the construction of new facilities. New instruments for attracting financing are needed, considering the high capital intensity and significant time costs for the construction of energy infrastructure. At that moment, discussions are underway to

introduce the latest technologies in the energy distribution model.

Nowadays, various start-ups, such as Grid+, Power Ledger, EnergoLabs, LO3 Energy and others, are developing blockchain-based applications for the electric power industry in many countries around the world. The reason for so much attention to the technology of the distributed register in the electric power industry is the prospects of optimizing not only the process of trading in power capacities, but also the process of energy generation and distribution of power capacities.

Blockchain technology is attractive due to some of its special properties. First of all, it allows you to automate the processes of accounting and some manual operations, which reduces transaction costs. Secondly, it allows you to save previously recorded information in its original form, which increases the trust of participants to the system and helps to track the conducted transactions.

The purpose of this paper can be described as the development of the most efficient model for conducting ICO (Initial Coin Offering) of a system of blockchain-options for the supply of electric power.

Dialectic methods, methods of system analysis, general scientific empirical methods were used to study the object of scientific research and achieve its goal: comparison, analysis, synthesis, method of scientific abstraction. Methods of retrospective, current and prospective analysis and synthesis of theoretical and practical material were used to systematize the data obtained. The methods used together made it possible to ensure the reliability of the research conducted and the validity of the conclusions drawn.

The material has been prepared with the results of studies carried out at the expense of funds provided under the grant of the Bank Santander.

II. RELATED WORK

The possibility of using blockchain in the electric power industry was considered by some authors, in particular, Mamontova M.U. drew attention to simplifying the interaction between generating companies and their customers through smart contracts [2]; Bogdanova E.D. and Valieva D.G. made a review of the use of distributed registry technology or the potential interest from companies in the energy sector [3]. Such big analytical companies like PwC [4] and Deloitte [5] carried out research in the field of blockchain in the energy sector and noted that the technology would help reduce the costs associated with user interaction with the audit; which will make transactions more transparent and help to establish fair prices.

According to research by GTM over the past year, blockchain start-ups attracted approximately \$ 300 million to the energy sector through the ICO [6]. Most projects are not currently implemented in practice, although they have identified in their studies a great potential for using the new technology. As part of the work, the documentation was reviewed Greeneum, Suncontract, Grid+, ImpactPPA, Power Ledger, EnergoLabs, LO3 Energy, Enerchain, WePower.

The introduction of modern financial technologies into the electric power industry is quite relevant and popular today on the market. For the purposes of this study, we analysed the works of a number of foreign authors. Thus, our team utilized the article "Security and Privacy in Decentralized Energy Trading through Multi-Signatures, Blockchain and Anonymous Messaging Streams", Nurzhan Zhumabekuly Aitzhan, Davor Svetinovic in which the authors discuss the problem of securing transactions in decentralized energy trading. They also proposed a concept for a decentralized energy trading system using blockchain technology. [7] Some provisions of this concept are reflected in the model developed by us.

You can also find traces of the following scientific works in the paper "Blockchain technology in the chemical industry: Machine-to-machine electricity market" by J. Sikorski, J. Haughton, M. Kraft (an example where blockchain is employed to establish a M2M electricity market in the context of the chemical industry) [8]; "A blockchain-based smart grid: towards sustainable local energy markets" by E. Mengelkamp, B. Notheisen, C. Beer, D. Dauer, C. Weinhardt (an example of decentralized market platform for trading local energy generation) [9]; "Industrial Blockchain Platforms: An Exercise in Use Case Development in the Energy Industry" by J. Mattila, T. Seppälä, C. Naucler, R. Stahl et al. (a case for autonomous machine-to-machine transactions of electricity in a housing society environment) [10].

During the process of finding technological solutions our team has used an experience of well known experts and communities, who developed smart contracts [11]. An external source of code and technological pattern was taken in order to construct and plan the synchronization system [12].

It was suggested that a decentralized platform could be built based on the analysis of the Russian electricity market, integrating the advantages of options and ICO, to form a

model in which both future consumers and private investors could participate in financing the construction of new facilities and granting access to cheaper energy.

III. RESEARCH OF ACTIVE PROJECTS, WHICH UTILIZE BLOCKCHAIN IN ENERGY SECTOR

The study of existing energy blockchain projects allowed us to identify the two most successful investment strategies - Grid+ and WePower solutions. The key factors were: signed agreements with energy companies, the availability of tested technological solutions and active marketing policy. However, it is still difficult to talk about the effectiveness of the projects reviewed in terms of comparing costs and financial results, since they are not fully implemented at the moment. Nevertheless, theoretically the benefits to suppliers and consumers are undeniable. It is established that the liquidation of intermediaries and P2P calculations will inevitably lead to a reduction in energy prices in a highly competitive environment.

One of the main advantages of the blockchain model is that all energy can be accurately recorded for certain suppliers and consumers. The control of the distributed and consumed energy makes it possible to ensure its optimal use. At the same time, a simplified distribution process, where consumers directly interact with producers, facilitates the free formation of prices in the market. On the one hand, a decrease in transaction costs will lead to a drop in prices, but on the other hand, the opposite effect is possible. For example, in the case of significant differentiation in producer costs, the price will not necessarily tend to average or minimize. As a result of the insufficient maturity of the digital market and the possibility of consumers sticking to the old redistribution system, the only loyal consumers will remain on decentralized sites. So, depending on the structure and volume of the digital energy market, the prices of the traded commodity will fluctuate.

The usefulness of derivatives in energy supply depends not only on the general provisions of the pricing mechanism, but also on the specifics of the market. Based on the fact that the option is intended to solve a well-defined set of tasks, you must be fully aware of the energy platform. Technological features in some cases can provide price stability and minimize the risks that lead to a lack of a need for options. Assigning certain rights and obligations to the token will inevitably lead to a restriction of its functionality. By itself, the integration of ICO and options is not the goal, it is only one of the possible tools for optimizing the interaction of market participants. Therefore, it is impossible to consider the methods of financing a project without an accurate idea of the specifics of the decentralized system.

Today tokens produced during the ICO can perform a variety of functions - payment for goods, works, services provided by the company; various discounts and bonuses; means of payment; system access keys; confirmation of ownership of an asset or a normal means of attracting financing and etc. For example, a token is the accounting

unit of a system whose properties depend solely on how the community will use it. Tokens can be secured with property and have a constant inflation rate. The company can also provide some monetary policy with it. The number of tokens produced by projects is also not limited, which allows creating different models that meet the requirements of the platform.

The following conclusions were made while analyzing existing energy projects on the market, regarding the efficiency of their use:

A. Projects with one token: WePower, Suncontract, Greeneum

It was possible to establish that using tokenization in these projects does not exclude speculation on the exchange rates of the coins on the platforms. The offer of the token will depend wholly and entirely on the number of users providing and purchasing energy. The more suppliers will be there, the more coins and the lower the price in conditions of weak demand. However, in the case of a very limited supply and high demand, attempts to dishonest fraud in the market cannot be ruled out, when profits will be used by suppliers for speculative purposes. In this case, the transformation of the token into a contract (option) would make sense. Nevertheless, it cannot be asserted that the use of derivative financial instruments will not exacerbate the situation in the long term and will not lead to even greater volatility.

B. Projects with two tokens: Grid+, Power Ledger, EnergoLabs, ImpactPPA, Exergy.energy

The model with two tokens was recognized as the most effective, based on the analysis of the decentralized companies providing energy distribution and trading services, which allows to distinguish between trading on the exchange (coin / national currency) and intra-system trade (coin / energy). This approach excludes the possibility of speculation, because each platform user can exchange a token only for energy. At the same time, minimization of risks of the parties in models with two tokens occurs due to the formation of a reserve fund, which can be used in the event of an emergency: a sharp increase in the costs of production of a supplier or a sudden drop in the solvency of consumers. The difference between the price of supply and demand will be offset by the resources accumulated by the system.

C. Models which do not imply tokenization of energy: Enerchain

The site allows to trade in gas, solar, wind and other energy without intermediaries. The expediency of the option in this case will depend on a number of factors, but the main one is the technological features of the platform. The combination of a market mechanism for pricing and the limited choice of consumers when pooling energy suppliers can lead to undesirable consequences. In fact, there will be a situation when agents of the offer will be able to "plan ahead" the profit received from the exercise of the option.

The solution of this problem can lie in the plane of improving the technological features of the platform.

IV. IMPLEMENTATION OF AGENTS` INTERRELATIONS MECHANISMS DURING THE WORK OF THE PLATFORM

The results of the analysis of the effectiveness of various option models, as well as the formed view of the appropriateness of holding options for the supply of power capacities, made it possible to draw up a model for the interaction of the parties. The proposed solution combines the advantages of previously analyzed projects and is designed to minimize the likelihood of instability in the energy market. Given the importance of the investment component, it was decided to form a blockchain model with three types of tokens:

1. An external token (ET), which is a unit of account. It is traded on a crypto exchange and can be exchanged for an internal or investment token of a vendor.
2. Internal energy tokens of individual suppliers (T1, T2, ..., Tn), which is necessary for energy trading, call and put options. Each internal token is provided with a certain amount of generated energy, which allows investors to guarantee the right to receive it in the future. At the same time, the token which identifies the energy will serve as a marker confirming the fact of the transaction.
3. The investment token (I1, I2, ..., In) - accounting unit for attraction of financing.

The conclusion has been drawn based on the results of the analysis. It is necessary to develop a qualitatively new model of energy tokenization. It is needed to preserve the advantages of using derivative tools when integrating with blockchain systems. The expediency of options in the model is mainly determined by the possibility of coordinating the interaction of market participants, rather than purely financial provisions. The relationship of rights and obligations, guarantees and market freedoms play a decisive role.

During the development of the decentralized platform model, special attention was paid to the risk-oriented approach. The risk assessment made it possible to identify the following possible sources of risk: legal regulation, investment attractiveness of the project, participation of third parties, hackers and the human factor. At the same time, risk factors were justified: the volatility of the crypto-currency, the electricity costs for the PoW protocol, the predominance of the role of large investors, the protection of personal data.

A. Market clearing mechanism

The general scheme of the proposed platform consists of several key elements. At the initial stage, the supplier receives a certain number of internal tokens T1, T2, ..., Tn, which is equal to the amount of energy generated by him, which he is ready to send to the system. These tokens are placed on the internal trading platform of the system, paired with an external token. That is, in order to acquire energy, the consumer must first purchase an external platform token on the common exchange, and then exchange it for the energy / internal token of the particular company (Fig. 1).

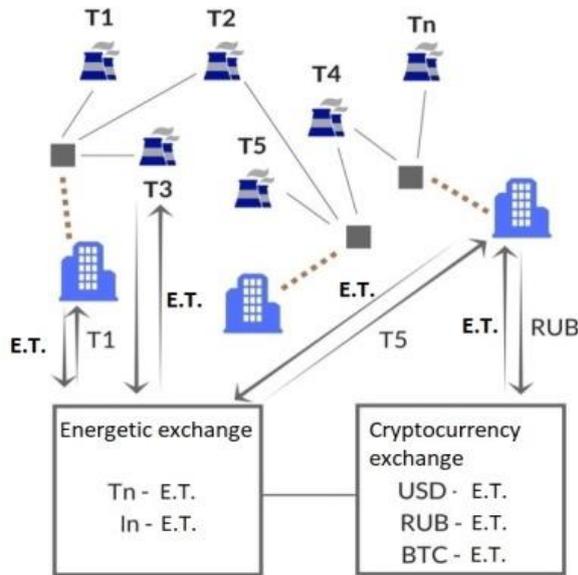


Fig. 1 Diagram of movement of external and internal tokens on the energy platform

In this case, clearing mechanism is based only on decentralized P2P interaction. Buyers can choose on the domestic exchange any token with a smart contract included in it. This coin has fixed the price and date and volume of delivery, established by the seller. The moment of buying the energy token on the domestic exchange will signify the beginning of the transaction on the blockchain option between the buyer and the energy supplier.

In addition, there is possible a following situation: when the energy buyer sets the terms of the smart contract and publishes it on the internal exchange, it can be prescribed to any generating company, which bought it.

The concept of the Waves project was taken as the basis of the internal stock exchange platform. The Waves developers created it as a decentralized exchange of tokens, sometimes abbreviated as DEX (Decentralized Exchange), due to the advantages of this type of system. DEX does not create limits for any transactions. After the process of buying cryptocurrency, it appears in your wallet instantly. Also, the exchange charges a very small fixed commission for each order. Transaction security does not cause fear among users, because the funds are stored in your wallet, not in the exchange. There is also no risk of losing money, as it may happen in a centralized exchange, but users still receive high bandwidth transaction channels and a centralized exchange node in order to maximize transaction speed.

Decentralized cryptocurrencies are traded on all centralized exchanges. Bitcoin and other cryptocurrencies are the ideal goods to steal. Transactions are not reversible; transfers are difficult to trace. Centralized exchange of cryptocurrencies is not sustainable in the long term, since the cost of a successful attack can spoil any successful exchange.

On the other hand, while there is a very high transaction capacity, the site will not be able to have a convenient solution for a fully decentralized trade. The inclusion of all

trading data in the blockchain is feasible, but then you have to deal with the delay of the lock and the set of data that must be synchronized across all nodes.

The solution can be simple - a decentralized block system, with a centralized order of checking the blocks. That is, there is a server that corresponds to incoming orders, but does not have access to resources. There is no chance that the user will be able to lose money in this setup, since he has full control over the funds. When the server finds the corresponding pair, it initiates the transfer to the block that moves the funds. Correctness of compliance is verified in a decentralized manner, and no funds can be transferred if orders are not verified (Fig. 2).

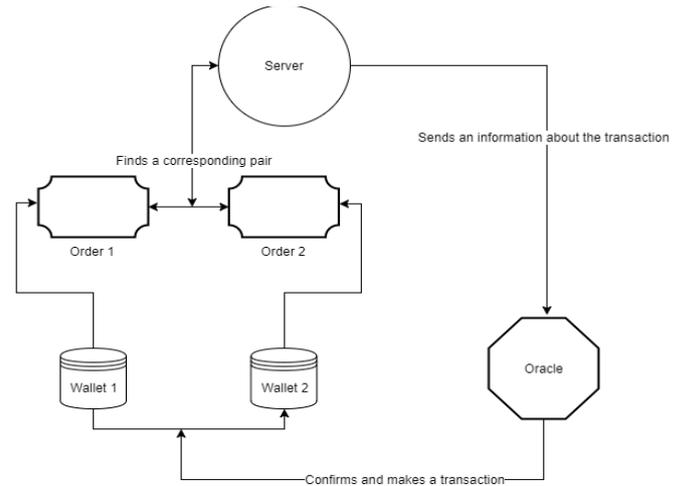


Fig. 2 Scheme of transactions inside the exchange

In accordance with the developed model, each of the energy companies or the directly generating stations, gets the opportunity to issue their own tokens in order to attract financing for the construction of new facilities or the updating of the existing ones. The exchange of each of the tokens I_1, I_2, \dots, I_n also occurs on the internal power exchange platform using the mechanism, similar to the process of buying an internal token. Consumers, in turn, get tokens by investing in the project. Those tokens are supported by energy, which they can exchange in the future for internal tokens T_1, T_2, \dots, T_n with discount by making a real option contract with the supplier, or exchange tokens for external coins (Fig. 3).

Often, energy companies, especially from the small and medium-sized business segment, find it difficult to attract investment to build or improve the electricity distribution infrastructure. Therefore, the platform envisages the creation of separate contracts to attract funds at the request of specific organizations. Thus, local residents in small settlements will have the opportunity to pay for the construction of power plants and power lines at their own expense or by third-party investors. The implementation of transactions can be made through the trade balance, with already available tokens of the particular company, bought at the rate for the E.T.

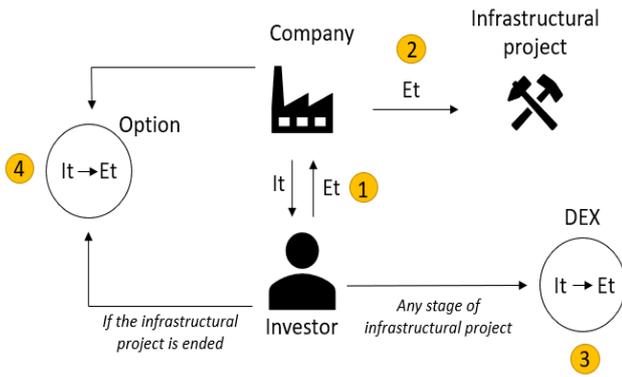


Fig. 3 The system of attracting investment – a general view

In the developed model options based on smart contracts allow you to build a clear scheme of interaction between the parties, delineating the rights and obligations. With the smooth operation of the system and full automation, it becomes possible to use options as tools to ensure the predictability of the energy market. All the transactions made at any time can be recorded on the platform, which means that it will be possible to track daily, seasonal and other fluctuations, providing information to suppliers for the future planning of power generation. The developed model of the decentralized platform seems to be the most optimal taking into account the requirements of the market and the level of development of modern technologies.

From a technological point of view, the proposed model for interaction between counterparties of a decentralized energy area includes the following participants:

1. The token creator, who creates a special token to ensure that the above-described smart contracts and the system as a whole are working optimally. Generating two contracts requires significant resources in Gas units on the Ethereum platform. As a result, the creator will be forced to absorb part of the costs when using smart contracts.
2. The Seller of the option is the user who enters the ERC-20 tokens in the amount indicated in the contract price. The option maker pays a commission.
3. Option buyer - the user who buys an option contract in exchange for the price set by the seller in the ERC-20 token through a protocol on the platform.

It is assumed that the option contract will give the right to receive special (ERC-20) tokens in exchange for the external ERC-20 tokens during the period before the contract expiration date at the strike price. A transaction using this right is called an "executive" transaction. A certain trusted third party that provides data on the current market condition in real time can be replaced by a platform administrator or by mutual agreement of its participants.

Blockchain allows to simplify and make more flexible the clearing system in the market of electricity sales. The decentralized exchange is the place of "meeting" of sellers and buyers, where everyone can interact with each other to find the best deal possible. The choice of price will help to make the transaction ledger on the blockchain, which will show the price trends.

B. Block-options' payment system

Two types of tokens will be used in order to conduct options - internal (T1, T2, ..., Tn) and external (Fig. 4). Call option, where the basic asset is electricity, is formed by the supplier. The consumer receives the right to supply energy at the specified time at a certain strike price after payment of the premium for the option. If the market price exceeds the exercise price at the moment of execution, the consumer realizes his right by paying the supplier a stipulated amount in external tokens. External tokens received by the supplier can be immediately sold on the exchange for other currency.

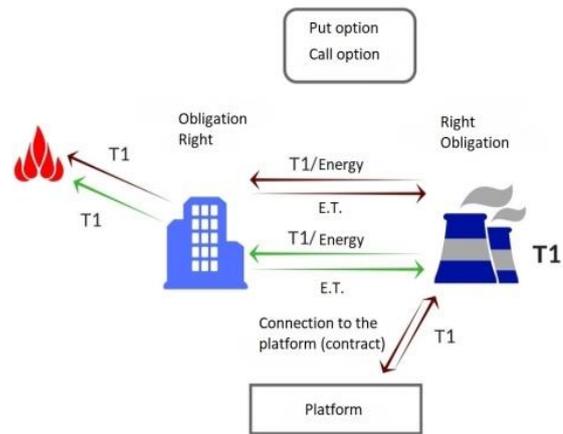


Fig. 4 Scheme of holding block of options for the supply of power capacity

The order of put option will be mirrored from the call option. In view of the need to formulate precise conditions for smart contracts, the flow of energy must also be accompanied by tokens. By connecting to the platform and providing information about the capacity of the power plant, the supplier receives a certain number of energy tokens T1, T2, ... Tn, which can also be used to participate in the put option. Each energy consumer, in turn, is entitled to form put option, indicating the desired date of energy supply and the amount for which he undertakes to purchase it. In order to confirm participation in the put option the provider transfers to the consumer account a premium in form of external tokens. When the option date is reached, the supplier decides whether or not to exercise his right. Accordingly, if the execution price is higher than the market price, the supplier confirms the transaction by transferring energy tokens identifying the energy to the consumer's account. Once this condition is met, the smart contract writes off the exercise price from the consumer's account. The internal energy token after the transaction is burned, confirming the fact of sending energy.

At the same time, the platform will be able to implement the bull call spread option, which reflects the strategy when an agent buys call options for a certain price and at the same time sells a call option for the same product, but at a higher price. It is used when an agent expects an increase in the price of goods. Additional add-ons to the existing system and a special user interface are needed to support the operation of such a mechanism on the basis of smart contracts. In addition, such a contract will need to control

the lower and higher price limits in order to activate itself if the latter is reached. To do this, it periodically updates the data from the market base of commodity prices (Fig. 5).

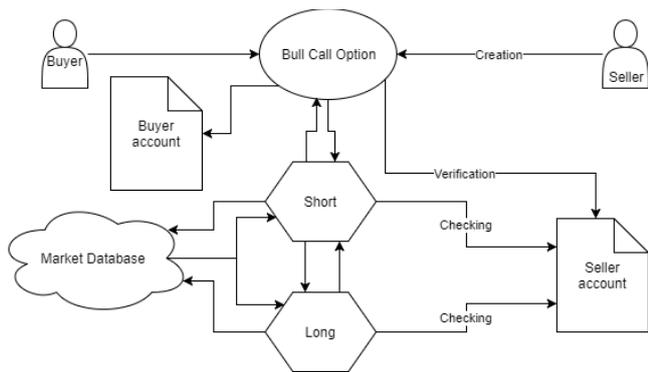


Fig. 5 Call spread option - detailed scheme

To determine this problem, it is necessary to raise the question: when and where is the code for the smart contract works? There are two different circumstances. Each smart contract is first performed by one or several miners, nodes that repeatedly offer new blocks to the blockchain. When the miner creates a block, he selects a sequence of user requests and performs associated contract codes for each Ethereum transaction in sequence, converting the old contract state to a new one. Then, it records both the sequence of transactions and the new state in the block, and suggests it for inclusion in the blockchain.

Later, when the block is added, each smart contract is re-executed by validators: nodes that restore (and check) the current state of the block. When the validator acquires each subsequent block, it repeats each transaction to verify that the initial and final states of the block match each other. Each miner checks the blocks proposed by other participants, and the senior unit is checked by newly connected miners or clients requesting the status of the contract. Code execution for validation significantly exceeds the execution of the code for mining [13].

The speed and coherence of the platform is very important for the acquisition of popularity among the customers of the market. Each smart contract must periodically update its state and check itself with the database. Unfortunately, the Ethereum Foundation, on which the developed platform is based does not have the tools to keep the smart contract network constantly updated, and attempts to add scheduled events to each element will only increase the server load and weight of the user interface. On the other hand, there is an Ethereum Alarm Clock system that is capable of this, but its work is based on the community of the network, and therefore it is not suitable for solving the tasks of the proposed platform, since it does not provide accurate data for the activation of smart contracts.

To solve this problem, we need a separate mechanism for updating contracts (Fig. 6). Creating a special smart contract that will connect all the other operating system elements to

the main server is one of the solutions. Of course, he will consume Gas, which will have to be compensated by commissions for the platform participants. Such server will help to regularly update price information and assist smart option contracts throughout their work.

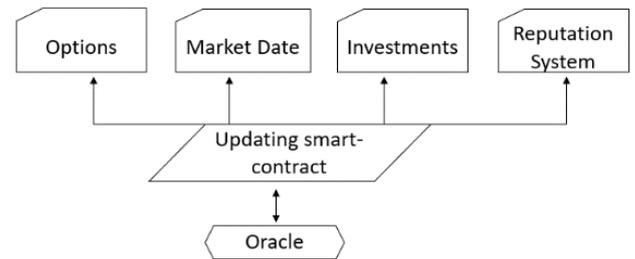


Fig. 6 Description of the work of the smart contract for updating the platform

However, Ethereum is already working on a full-fledged system for registering updates and accounting of time in smart contracts, which will avoid the listed difficulties associated with the development of its own solution.

To maintain the workability of the platform, it needs a source of finance. The main contender for this function is the system of commissions to work for the creation and execution of smart contracts at the Ethereum Foundation. Each smart contract requires a certain payment in the equivalent of Gas. This is the internal crypto-currency of Ethereum, through which the cost of production and placement of the contract is on the platform is paid. Therefore, when placing such a tool in our system, a certain amount of money will be written off in the form of tokens.

Many projects to date show that programming and the use of financial instruments on Ethereum is quite possible. However, while the underdevelopment of the structure makes the creation of such projects, cumbersome and complex, those solutions can be used in the market and give companies the necessary and unique functionality that makes them more successful than their competitors.

For the proper work of financial instruments, it is necessary to establish a communication with external sources of data on the market and counterparties. Unfortunately, at this moment, there are no reliable suppliers of such information such as Bloomberg on blockchain technology.

However, there is a project called Ethereum Price Feed, under the leadership of Roman Mandeleile, which can be used as a third-party source for smart contracts. On the other hand, this platform is updated. To solve this problem, several options were considered, including the one described above, but in the end of the simplest answers was chosen to use data from open sources such as coinmarket.com and other websites analyzing quotes of crypto-currencies online. Also, to get the prices of commodities, you can use Bloomberg or other local sources, depending on the location and methods of using the platform. These data will be used

as the main indicators for smart contracts and, accordingly, users, too.

Sources of price data can work much better in closed blockchain systems. As with any other similar platforms, the market and industry indicators can be defined and tailored to industry standards. Providers will not be able to. Platform participants can also enter their data on the state of markets and if they are confirmed by other agents, their use becomes possible within the framework of a particular project.

C. *Interaction risks' decrease*

To implement a full-fledged trading platform requires a base in the form of a website or an application for a user interaction between each other and the blockchain system as a whole. In addition, one cannot force people to use their wallets, as this will involve third parties. That's why every user of this platform will need to open his / her own trade balance directly on the platform. The numbers of these trade balances will be used by the system as input for settlement through smart contracts.

Once one of the users decides to open such an account, he will be asked to put on a certain amount of money. This is designed to protect participants 'wallets, since a smart contract is the right to manipulate users' funds on their trade balances. Subsequently, such a contract can withdraw tokens directly from accounts of counterparties. In the future, to ensure the protection against intruders, smart contracts can be connected to the system for determining and recording the reputation of participants. Contracts also provide access only to persons who have made a transaction on their basis, which makes the intervention of the third party unlikely.

A special section will be opened on the website and in the project in order to promote and regulate the market. It will allow users to communicate with each other and influence the fate of the project as a whole and the decisions of local companies in particular.

However, this approach seems ineffective, since most users can ignore the platform's capabilities, leaving room for intruders. The special reputation system will follow to control the regulation, which is also tied to contracts and trade balances of market participants.

Another practical solution will be the creation of a reputation accounting system [14]. Systems of accounting for the reputation of users are not new on the blockchain market. There are many similar projects based on Ethereum, such as Augur and Gnosis. They are based on the value of money and hold important information. Participants in these projects receive more tokens if they have a good reputation among other users.

The main part of this system will be reputation points. In essence, they are similar to the credit score. When a person takes out a loan for the first time, he has no history, so at the beginning each agent will be given a certain "starter" rating, which will be slightly above zero.

Each creator or buyer of options has an open profile in the community forum, the data from which are displayed in the authorship of the contract. Thus, any participant will be able

to put this or that rating to another user and this data will be stored in the blockchain network.

Of course, one should not forget about the possible abuse of the system, but this problem has already been partially resolved, since each new agent of the platform is obliged to make a minimum deposit for making transactions on the trading floor, which makes the creation of additional balances at least unprofitable.

Also, the system will take into account each user in order to make buyers, sellers and investors more significant if they have a good reputation among the community. This function will prevent Sybil type attacks on the blockchain network.

Another side of the reputation system will be the valuation of the energetic companies. The risk rating will be applied to them, and the profitability of the company. Also there will be a data on payments on investment shares and the total profitability ratio.

For example, the reputation area was taken from 0 to 1000 rating units, where 0 is the worst and 1000 is the best. Weight in this model was not specified in order to simplify the presentation.

Depending on the behavior of each individual agent, the community will be able to identify both conscientious and unfair users.

Such a system will allow rewarding the participants with the best reputation, giving them the opportunity for further development and opening up new market sectors. Over time, this will create an environment in which such platform users can, even if it is a local supplier or consumer.

The system also plans to take into account the total number of transactions and contracts. Continuing to use the platform, the reputation of the actors will be constantly updated, giving a chance to agents with a bad reputation to correct themselves and showing the trend and the history of everyone.

V. CONCLUSION

Thus, the following main results of the study were obtained:

- Evaluation of the effectiveness of blockchain technologies in the electric power supply industry;
- The technological and market justification for the release of several interconnected tokens on the platform;
- Proposals were made to establish relationships between contractors based on the platform;
- The method of implementation on the platform of specialized smart contracts.

The practical significance of the research work is to compare methods of applying traditional derivative financial instruments and blockchain options in the energy supply industry with the subsequent creation of the best model for the development in this sector.

Summarizing the above, based on the results of the work, the most effective model for Initial Coin Offering was developed for the system of blockchain options for the supply of electric power. Also, the principles for the functioning of

the decentralized site were formulated during the research. Realization of the proposals will allow, on the one hand, to optimize the process of trade in power capacities, and on the other hand, will lead to an increase in the efficiency of the generation and distribution of electric power.

APPENDIX

A list of terminology that was used in the paper:

Distributed ledger – A distributed ledger (shared ledger, or distributed ledger technology, DLT) is a consensus of replicated, shared, and synchronized digital data geographically spread across multiple sites, countries, or institutions. There is no central administrator or centralized data storage.

Blockchain – A blockchain, originally block chain, is a continuously growing list of records, called blocks, which are linked and secured using cryptography. Each block typically contains a cryptographic hash of the previous block, a timestamp and transaction data.

Smart contract – A smart contract is a computer protocol intended to digitally facilitate, verify, or enforce the negotiation or performance of a contract. Smart contracts allow the performance of credible transactions without third parties. These transactions are trackable and irreversible.

ICO (Initial Coin Offering) – An unregulated means by which funds are raised for a new cryptocurrency venture. ICO is used by startups to bypass the rigorous and regulated capital-raising process required by venture capitalists or banks. In an ICO campaign, a percentage of the cryptocurrency is sold to early backers of the project in exchange for legal tender or other cryptocurrencies (usually for Bitcoin).

Call option – A call option, often simply labeled a "call", is a financial contract between two parties, the buyer and the seller of this type of option. The buyer of the call option has the right, but not the obligation, to buy an agreed quantity of a particular commodity or financial instrument (the underlying) from the seller of the option at a certain time for a certain price (the strike price). The seller ("writer") is obligated to sell the commodity or financial instrument to the buyer if the buyer so decides. The buyer pays a fee (called a premium) for this right. The term "call" comes from the fact that the owner has the right to "call the stock away" from the seller.

Put option – In finance, a put or put option is a stock market device which gives the owner of a put the right, but not the obligation, to sell an asset (the underlying), at a specified price (the strike), by a predetermined date (the

expiry or maturity) to a given party (the seller of the put). The purchase of a put option is interpreted as a negative sentiment about the future value of the underlying stock. The term "put" comes from the fact that the owner has the right to "put up for sale" the stock or index.

REFERENCES

- [1] F.V. Veselov, A. Khokhlov, *Internet of Energy: how distributed energy will affect security, prices for electricity and ecological situation*, Russian version of Forbes, section "Business", October 18, 2017 [In Russian]
- [2] M.J. Mamontova, "Blockchain and opportunities of its implementation in energy", *Information technology in science, management, social sphere and medicine, Collected scientific proceedings of IV International scientific conference*, December 5-8 2017, Tomsk – pp. 417-419 [In Russian]
- [3] E.D. Bogdanova, L.G. Valieva, "Cryptocurrency and energy", *Problems, prospects and tendencies of innovative science development: collected articles of international academic and research conference*, 2017, pp. 63-67 [In Russian]
- [4] PwC, *Use cases for blockchain technology in Energy & Commodity trading – 2017*, [online] Available at: <https://www.pwc.com/gx/en/industries/assets/blockchain-technology-in-energy.pdf> [Accessed 2018]
- [5] Deloitte, *Blockchain applications in energy trading*, 2016, [online] Available at: <https://www2.deloitte.com/global/en/pages/energy-and-resources/articles/role-of-blockchain-in-the-energy-and-resources-industry.html> [Accessed 2018]
- [6] CCN, *Energy Sector Invests \$300 Million In Blockchain In Past Year*, [online] Available at: <https://www.ccn.com/energy-sector-invests-300-million-in-blockchain-in-past-year/> [Accessed 2018]
- [7] Aitzhan, N. Zhumabekuly, and D. Svetinovic, "Security and privacy in decentralized energy trading through multi-signatures, blockchain and anonymous messaging streams", *IEEE Transactions on Dependable and Secure Computing*, 2016
- [8] J. Sikorski, J. Haughton, and M. Kraft, "Blockchain technology in the chemical industry: Machine-to-machine electricity market", *Applied Energy* No. 195 (2017), pp. 234-246
- [9] E. Mengelkamp, et al., "A blockchain-based smart grid: towards sustainable local energy markets", *Computer Science-Research and Development* no. 33.1-2 (2018): 207-214
- [10] J. Mattila, et al. "Industrial blockchain platforms: An exercise in use case", *Development in the energy industry* no. 43. The Research Institute of the Finnish Economy, 2016
- [11] Anon, (n.d.). *Smart Contract Application Examples and Use Cases ...* [online] Available at: <https://www.draglet.com/blockchain-applications/smart-contracts/use-cases> [Accessed 2018].
- [12] Anon, (n.d.). *Smart Contract - A Fully Decentralized Oracle Network*. [online] Available at: <https://link.smartcontract.com> [Accessed 2018].
- [13] Arstechnica. *Bitcoin's insane energy consumption, explained* - [online]: URL:<https://arstechnica.com/tech-policy/2017/12/bitcoins-insane-energy-consumption-explained/>
- [14] M. Kindy, *Divine: A Blockchain Reputation System For Determining Good Market Actors* <https://medium.com/topl-blog/divine-a-blockchain-reputation-system-for-determining-good-market-actors-7c47a0308ae/> [Accessed 2018].

B2B Price Management using Price Waterfall Model and Business Intelligence solution

Krzysztof Senczyna
Czestochowa University of
Technology, ul. Dabrowskiego 69,
42-201 Czestochowa
Poland
Email: krzysztofsenczyna@gmail.com

Radek Němec
Faculty of Economics, VŠB –
Technical University of Ostrava,
Sokolská třída 33, 702 00 Ostrava,
Czech Republic
Email: radek.nemec@vsb.cz

□ **Abstract**—The price setting and negotiation process in the B2B field is a complex process that requires a solid methodology and usually also advanced IT tools to make the process as efficient as possible. The Price Waterfall model is a flexible tool that allows for making the final price determination and revenue creation task much more manageable. In this paper, we introduce a software solution which integrates functionalities of a standard Business Intelligence system with a methodology given by the idea of the Price Waterfall model. The tool is designed as a dedicated decision-making support tool, with a complex internal workflow that should be applied within the price and revenue management process, to induce profitability of the whole business through informed decisions.

I. INTRODUCTION

THE field of B2B interaction between companies puts more and more emphasis on the price negotiation processes because each company wishes to maximize profit margin from every business transaction. Therefore, the emphasis on building excellent decision-making support solutions, that use advanced data analysis methods as well, is getting stronger. The Digital Economy generates data as a basis for many solutions and models in the IT implementation area. For each organization equipped with the appropriate potential, the use of this data is an important factor for supporting decision-making processes.

The Digital Economy also changes the way the price is negotiated mainly in the B2B sector (but recently in the B2C as well). In this paper, we focus mainly on the B2B field, because the success in this field influences success factors of B2C interactions – fields are mutually beneficial [22]. Also, technically, the B2B can be viewed as a prerequisite regarding functional relationships between key players on the market.

The demand-and-supply law and the "invisible hand of the market" regulate the price. At the same time, however, as shown by Mc Kinsey's research, the 1% price increase generates 10-11% profit increase (*The power of 1%*) [1]. The success of the price negotiation process is, therefore, very important. However, as such, it is also very prone to a lack of accurate and timely information. Therefore, the execution of

the price negotiation process is often backed by dedicated tools and computing capabilities within the enterprise information system, to minimize risk of possible economic losses.

The research on price management issues encompasses many research topics, e.g., from analysis of trends in pricing systems, to importance of exactness in the pricing process, relationships between price, revenue management, and business performance, dynamic pricing computations and approaches, use of price optimization in various use cases, competitive price information in the revenue management, and last but not least, pricing frameworks in competitive industries [29], [30], [31], [32], [33], [34], [35], [36], [37].

Concerning the information systems that support the business, and its success in terms of information delivery timeliness and relevancy, the relationship with the execution of related business processes is very important [2]. A precondition of the decision-making process efficiency is a smooth and seamless adaptation to changing business conditions, through the use of decision-making support tools. Such tools should offer innovative as well as added-value functionalities, combining well-known business analytics methods with business reporting capabilities, using innovative approaches (innovations are important change drivers within organizations [3]). An intuitive decision-making process is sometimes mentioned as an alternative approach in the management field, especially in ambiguous or uncertain situations [10]. However, in most business critical situations, the use of sophisticated computerized decision-making support, like the Business Intelligence (BI) system, is a complete necessity [15], especially when it comes to the issue of mining large data sets.

Usually, major BI system solutions offer standard functionalities, and in numerous companies around the world, these tools are still the most widely used ones within the decision-making support [4]. Specialized business processes, like the price setting and negotiation management in the B2B field, usually require specific back-room algorithms and functions to be executed seamlessly, and in full accordance with business users' expectation: consumption of ready-to-

□ This paper was made with financial support of the European Social Fund within the project CZ.1.07/2.3.00/20.0296 and the Student Grant

Competition project SP2018/146 "Evaluation of comparison applications using cognitive analysis and the Data Envelopment Analysis method".

use outputs from the system, to make actual decisions more easily. Standard BI system functionalities are usually very broad, regarding their usage in specialized use cases, and certain calculations can be complex enough to employ different strategies.

Dedicated business-analytics and reporting-based solutions then come in mind to fill the functionality and output-interconnection gap, so that the whole decision-making system's environment moves closer to the idea of IT support of an intelligent enterprise. Such concept is mentioned in [5], and also in [19], in the context of real-time decision-making within an autonomous supply chain system. Cloud computing is commonly mentioned in connection with improvements of efficiency in the field of delivery of data analysis and reporting functions for business users. Cloud based BI systems, especially if the focus of the system is to interconnect various types of data source, offer many benefits, like cost efficiency, flexibility, and scalability, along with enhanced data sharing capabilities [21].

In this paper, we present an enterprise-grade solution that bears the characteristics of the above-mentioned system, and focuses on the fulfillment of requirements, that stem mainly in the B2B price management field. The solution leverages Price Waterfall model as a methodological background.

The further presented system leverages analysis of large datasets, to facilitate its main purpose – the price management process execution. In the system, there is the knowledge from the field of computation performance optimization applied – (sales) data vectorization approach, which is a key feature of the Price Waterfall model implementation (data vectorization features and benefits were studied e.g., in [27] or [28]).

A game-theoretic approach is also a promising approach, since it enables the use of dynamic and competitive price modelling patterns (in [26], there are benefits of data-driven competitive analysis approach mentioned, in terms of creating price setting system architecture). Although the game-theoretic approach has important features, The Price Waterfall methodology allows to implement tools for a full scale pricing process execution (from analysis to price and contract configuration and further management).

The paper is structured as follows. First, the price management background of the software and the Price Waterfall model the software are described. Secondly, the software solution is introduced, as a complex, BI-based, price configuration and management toolset. Finally, the functionalities of the CPQ (Configure, Price, and Quote, [14]) software solution, focus on leveraging the Price Waterfall model, are presented in more detail, using a sample dataset.

II. PRICE MANAGEMENT BACKGROUND AND THEORETICAL ASSUMPTIONS OF PRICE WATERFALL MODEL

In the world of digital economics, there will soon be no place for organizations that do not analyze their economic-activity and related data – it will simply lead to exclusion from the economic world. In contrast with that, and as quoted in [13, p. 14], the pricing is “a messy business” that varies greatly between industries and even different companies and as such, it is a subject of ongoing research.

There are other standards of customer behavior appearing in the B2B and B2C sectors. Within the B2B sector, the online shopping phenomenon is most visible, as research results in [6] confirm it – 93 % of B2B buyers prefer to buy on-line when they've already decided what to buy, and 93 % of B2B buyers prefer not to interact with Sales Rep. as their primary source of search for information. Other forecasts and behavior observations imply that, e.g., [7]:

- by 2018, more than half of large organizations globally will compete using advanced analytics and proprietary algorithms, disrupting entire industries;
- by 2018, 40 % of B2B digital commerce sites will use price optimization algorithms and configure/price/quote (CPQ) tools to dynamically calculate and deliver product pricing.

The most important decision-making process in the company's structure should be the process of determining the price. In the dynamic market situation era, where changes in prices and costs within enterprises happen continually, it is strategic to capture these change moments. It allows us to adjust and change prices to higher prices as late as possible and as soon as possible if such a possibility occurs.

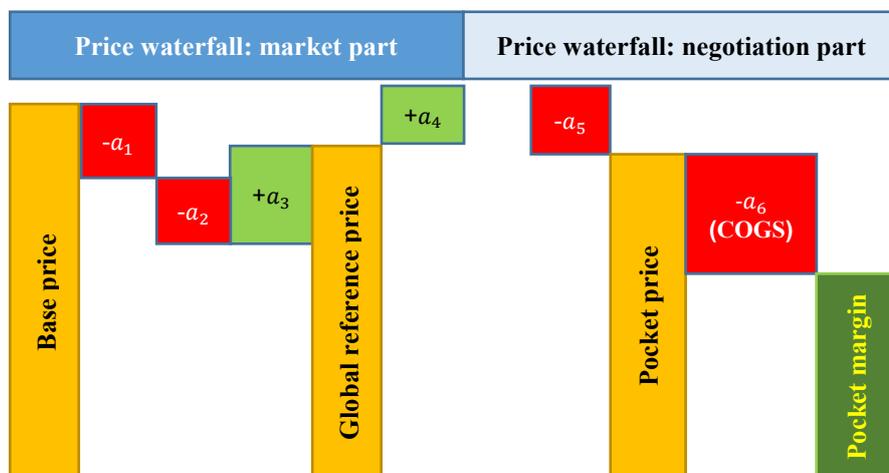


Fig. 1 Example of a Price Waterfall model visualization (*COGS* = *Cost of Goods Sold*, i.e., *standard product costs*)

Sometimes the fact of maintaining stable prices for a certain time is valuable and a highly desirable fact.

Market conditions press on the goods and services delivery so that the price is at the lowest possible level and the entire contract is more profitable for the customer. But on the other side, the whole transaction should be kept profitable also for the reseller. These two contrary motives push forward the necessity of implementing systems for an efficient determination of price in time, through a collection of historical data and by design of price trends, to be prepared for eventual price fluctuations and customer-side requests. The discussion on the use of the actual models and methodology, then comes to mind.

A. Price Waterfall model, price adjustments and Pocket Margin

As mentioned in [8], the idea of the Price Waterfall model (PWm) is a mapping in the form of successive degrees as in the cascade of factors affecting the determination of the price of a given contract.

The process starts from Basepoint, in which the Base price (Base price point) is set, based on historical data (i.e., certain verification of the price by the market is already available). In the next steps, the price in consecutive price points changes according to **adjustments** that may be positive or negative (given the nature of the actual adjustment). Calculation of *k*-th price point value is carried out using equation 1:

$$PricePoint_k = PricePoint_{k-1} + \sum_{i=1}^n \pm a_i, \quad (1)$$

where $\pm a_i$ represents *i*-th price adjustment out of total *n* adjustments set after the establishment of previous price point $PricePoint_{k-1}$. Each price point refers to a certain point in the price setting and profitability assessment process (fig. 1

depicts a sample price waterfall and resulting pocket margin). If the sum of price corrections is negative (i.e., if there are mainly negative adjustments $-a_i$), the contract is discounted.

Each price adjustment between price points refers to different contexts, e.g., product attributes, bundling rules, regional pricing rules, channel adjustments, standard discounts, or negotiated discounts, service and shipping charges, rebates, service costs, and finally the standard product costs.

Market configuration leads to an Invoice after the Discount point where there is already a price established, which appears on the customer invoice. After this point, the cost part appears in the waterfall, where all costs related to a given transaction are included.

The last price point is called **Pocket Margin (PM)**. Through the computation of the Pocket Margin value, as the last element of the waterfall cascade, at the end of the transaction (i.e., the end of the entire price setting process), an estimation of the profit from this transaction is done. This feature of the PWm, i.e., a prompt estimation of the profit from the transaction at the moment the transaction is created, seems to be the most important in the whole process. Below-zero Pocket-margin value indicates profit (price margin) leakage, e.g., too high discounts were awarded, or adjustments at a given stage have not been done correctly, etc.

This phenomena should be identified as soon as possible to stop the profit leakage so that the whole price management process is efficient [25]. An example of a helpful analytic visualization of PM data is shown in Fig. 2 – a Tree Map visualization using transaction data (within the further presented CPQ solution). The red color squares show the area where $PM < 0$, and it request a deeply analysis to get verify the reasons.

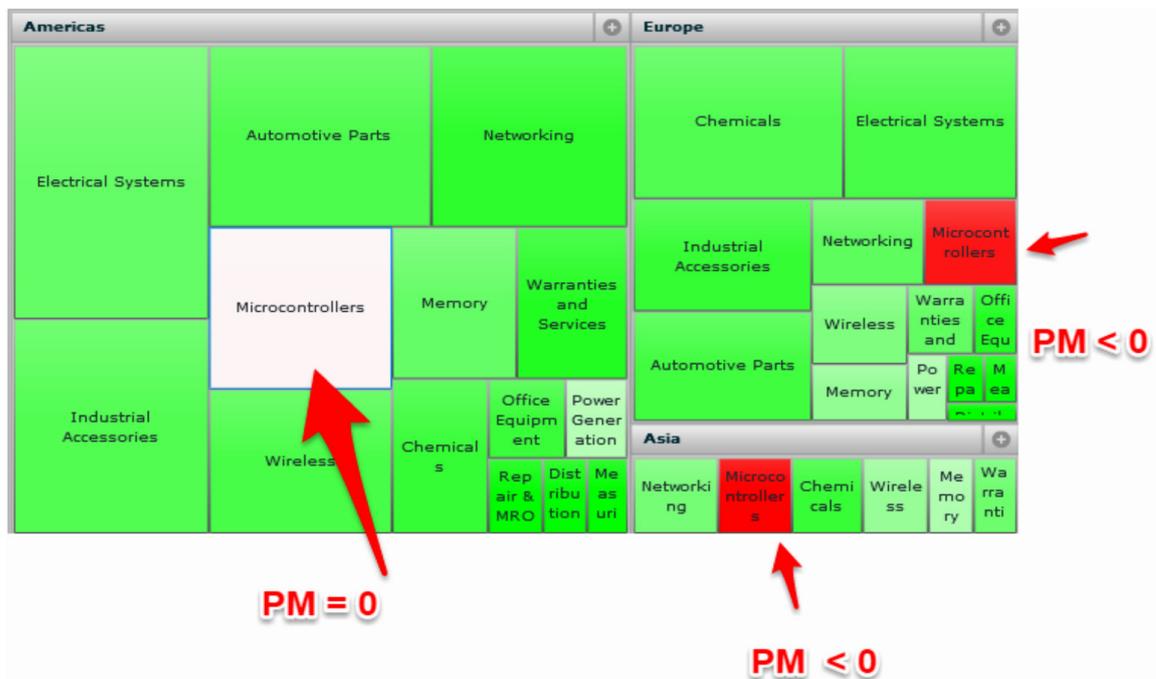


Fig. 2 Example of a Tree Map visualization used to analyze Pocket Margin values

Transaction data and the structure of the price waterfall cascade then allow for a calculation of the expected profit value. Depending on the defined pricing policies, the transaction will be prepared for the client after that, and they can then apply for additional discounts. Nevertheless, the limiting factor, i.e., the value of the Pocket margin is given for the transaction, and it should be kept greater than zero.

B. PWm and sales data vectorization

Implementation of the PWm-based price determination process (e.g., into a CPQ solution discussed in this work) ultimately makes possible the use of the strongest advantage of the PWm – **the data vectorization**¹.

For each transaction, data vectors **WF** of the price cascade is created. Anything in the sales process, i.e., the Customer, Sales Rep., Channel, etc. has its own defined **WF** vector for a given time range. In this way, it is possible to compare object's data by comparing the **WF** vectors of these objects, for example, $\mathbf{WF}_{\text{Channel (1,year 2017)}}$ vs $\mathbf{WF}_{\text{...Channel (2,year 2017)}}$, etc.

At any moment, the **WF** vector is specified, and this event enables tracking of changes for the entire company as it allows to compare relevant periods. Of course, the transaction data analysis without a pricing model can be done using tools like *QlikSense* or *PowerBI*. However, within the data analysis based on the PWm, the advantage in the form of data vectorization allows us to use of a more complex and multi-dimensional approach to the analysis of certain phenomena in sales activities.

The Big Data phenomenon [16], generally, and the analysis of very large data sets also plays an important role in the process of rich business insights creation [9]. Big Data is a source of many opportunities in multiple areas, like an increase in operational efficiency, creation of informed strategic decision, and also better customer service, etc. [17], [18], [20]. It is reported that in the commerce field, the use of Big Data analytics can lead to a 60% increase in operating margin [11]. So the use of Big Data sources within the price management process should be viewed as a valuable source of insights as well. Certain external data, like data from users' interaction within social media, can contain information that may lead e.g., to definition of additional types of price adjustments.

Other modern and mostly unstructured data sources, like customer expectations expressed as opinions within social media posts and comments, or video blogs and voice recordings may contain such information². Through application of well-known unstructured data analysis methods, like sentiment analysis and natural language recognition and processing in general, the way to the overall improvement of the price and revenue management is already open.

¹ Fluid use of data vectorization within the data analysis process is a computationally non-trivial task, but may lead to richer insights.

III. PRESENTATION OF THE CPQ PRICE MANAGEMENT SOLUTION

The further presented CPQ solution leverages the PWm methodology and enables interconnection of data from various enterprise sales processing systems, to analyze data and presents outputs that allow for automatic price adjustments and very quick decisions. However, only the use of structured data sources is currently implemented in the system (fluent processing and analysis of unstructured data is one of future milestones). The system includes standard as well as more advanced BI-system based functions (among other), which are a necessity today. The software solution is primarily intended to be used by standard as well as power users within the pricing process. User interface and back-room functions of the solution are programmed mainly using the Java programming language.

A. Main modules of the CPQ solution

The full scope of data processing includes three key modules: the module for the data analysis and visualization (Analyze), the pricing policy configuration module (Optimizer), and the contract creation module (Execute). The results obtained from the contract creation module are returned to the Analyze module, thus closing the data flow cycle (fig. 3). The trend of decision-support systems integration, in the field of complex Management Information Systems deployment, rather than creating isolated systems (as mentioned in [12]), is fully respected in case of the presented software solution (modules cooperate with each other).

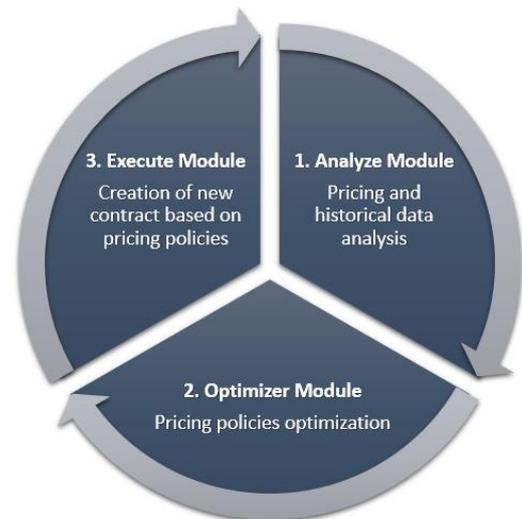


Fig. 3 Data flow within the CPQ solution's environment

Main tasks of the Analyze module:

- Identify price, margin and profit opportunities for any part of the company's business, utilizing transaction data;

² Some of these may be, or already are included within certain proprietary price setting rules, but without proper tools, they can be used as rather vaguely defined indirect effects.

- analysis of a number of business metrics;
- identification of specific areas of margin leakage (i.e., with $PM < 0$);
- visualization of profit opportunities and contributing factors with a possibility of sharing the results with team members for higher productivity;
- explanation of how revenue or margin changed from one period to the next regarding the price, volume, mix, win/loss, cost, and exchange rate effects.

Main tasks of the Optimize module:

- Combination of all relevant internal and external data needed for setting the price (various product costs, competitive and market information, past pricing process performance, etc.), and setting the prices in a single rule-based system combining both data classes;
- generation of massive amounts of prices using configurable rules and strategies, and easy management of price lists, thus enabling efficient mass price changes;
- tracking of price changes (workflow is recorded);
- integration with downstream systems for pushing and publishing prices.

Main tasks of the Execute module:

- Evaluation and enforcement of pricing strategy on every deal;
- modelling each deal (quote, contract, etc.) for profitability estimation, with adherence to pricing strategy, etc.;
- automatic routing of deal to appropriate approvers based on each deal's characteristics (workflow enabled), thus enabling both simple and very complex deal management, including mass price changes.

Fig. 4 shows standard reporting and visualization charts within the CPQ solution's Analyze module.

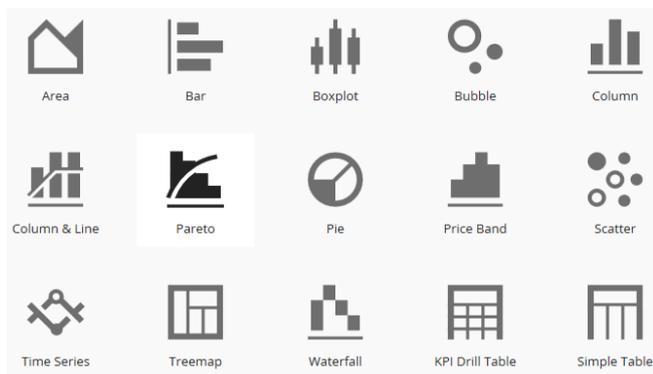


Fig. 4 Standard charts within the Analyze module in the CPQ solution.

Some functions are also available in BI tools available on the market (e.g., the bar chart, box plot visualization, scatter plot, etc.), as well as visualization functions that are strictly related to the PWm concept – Waterfall and Price Band. Each function supports the idea of decision-making automation,

within the price configuration and management process, i.e., each one can be used to visualize certain aspects within the process, with the possibility of viewing special features of the process using both special functions mentioned above.

B. Example of Price Waterfall visualization using sample dataset

In fig. 5, there is a sample of the Price Waterfall visualization (using the “Waterfall” function) presented, generated using a sample dataset. The test dataset was used due to security reasons because it was not possible to showcase live company's data in the paper. The dataset is a standard multidimensional dataset for OLAP-based data analysis, with dimensions like Products, Country, and a fact table, with columns containing values of actual price points and adjustments for a given transaction time.

The use of vectorized sales data for data analysis also allows us to use 3-D charts. This way of data visualization is possible because each transaction has its price data vector **WF** set. 3-D visualization of the price waterfall is one of possible future enrichments that could be implemented within the presented CPQ solution. Principally, the 3-D visualization of the price waterfall allows for a deeper analysis of the price establishment and profitability management process (output combines benefits of popular charts like tree map and bubble chart). Currently, the tool's UI lacks such functionality, so an example of such visualization was elaborated using Microsoft Excel's charting functions – result can be seen in fig. 6.

As seen in the fig. 6, in the upper right corner, there are companies grouped for which the Pocket margin value is high – i.e., these are very profitable customers. The situation in the bottom left corner should be analyzed because the Pocket margin value is low and even less than zero. There is possibly a transaction in this area, for which the Base price is high, and yet Pocket margin is close to zero or negative (Invoice price value is also shown there). So losses have been incurred and detected with a possibility of obtaining more precise insights (the reason for these decisions should be analyzed as soon as possible). After including the third viewpoint (Base Price in this case), the visualization of the Price Waterfall shows more promise than the standard 2-D variant, especially if there will be more such outputs included in a complex dashboard (i.e., more business performance aspects could be studied at once).

C. Known limits of the solution and future outlook

Known limitations of the software are that it focuses mainly on the CPQ problem, i.e., the process of ad-hoc setting and price management within the B2B relationship. Currently, there are no functions that would possibly allow for a prediction of future prices (or price adjustments) as well as the revenue of the company, or even relationships on the B2C (C2B) level. The inclusion of B2C-related capabilities are, however, more important in the case of relationships with non-enterprise customers (which is actually not the case of the presented solution). In this field, however, the analysis of customer-based knowledge about our products' aspects would be essential [23].

Capability of predicting future states of important business aspects, using machine learning or even deep learning methods, to understand text, emotions and, generally, the unstructured data, are becoming a crucial functionality [18] (one instance of a smart quotation system is presented in [24]). The fact that not only large companies recently started to recognize the importance of Big Data sources implies that the vast amounts of multi-purpose data are very tempting. As it was mentioned above, it is important and also highly relevant to the price and also revenue management process. E.g., events' features and descriptions that may contain hints for future events, people's interests and their development that may induce future changes in demand, etc. Mining also such data source is in the future plans of the presented CPQ solution's development.

IV. CONCLUSION

The CPQ tools market will grow significantly in the future. It will be driven by the need for advanced price management solutions, with powerful predictive data analysis capabilities that, among others, will help to optimize such a crucial decision-making process, as is the optimal price creation process. The integration of such functionalities and capabilities in an integrated BI system environment will also be a necessity, given the amount of and data interdependencies and necessary visualization options, for the presentation of results in a way that allows for a continual improvement of the company's business.

Next steps in the development of the presented CPQ solution will be the expansion of capabilities with the utilization of AI and machine (deep) learning algorithms as well as interconnection with insight-richer data sources (Big Data sources). It would lead to the creation of more advanced system that would allow for price optimization within the revenue management process (by the PWm methodology).

Future research on this topic will focus on the revenue optimization and related approaches, with a study of its possible efficient inclusion into novel functions. These new functions might allow for an optimized execution also of the revenue management process, within the interface of the software solution.

REFERENCES

[1] E. Maltby, "Raising Prices Pays Off for Some," *The Wall Street Journal*, 2010.

[2] S. Petter, W. H. DeLone, E. R. McLean, "Measuring information systems success: models, dimensions, measures, and interrelationships," *European Journal of Information Systems*, vol. 17, 2008, pp. 236-263.

[3] R. Némec, F. Zapletal, "The Perception of User Satisfaction in Context of Business Intelligence Systems' Success Assessment," *Proceedings of the IDIMT-2012: ICT Support for Complex Systems: 20th Interdisciplinary Information Management Talks*, 2012, pp. 203-211.

[4] M. Łobaziewicz, "The Role of ICT Solutions In the Intelligent Enterprise Business Activity," *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, vol. 8, 2016, pp. 1335-1340. <http://dx.doi.org/10.15439/2016F534>.

[5] R. Némec, "Assessment of query execution performance using selected Business Intelligence tools and experimental agile oriented data modeling approach," *Proceedings of the 2015 Federated Conference on Computer*

Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, vol. 5, 2015, pp. 1327-1333. <http://dx.doi.org/10.15439/2015F267>.

[6] CRM Magazine, "Death of a (B2B) Salesman?," digital version, 2015, url: [http://www.destinationcrm.com/Articles/Columns-Departments/Insight/Death-of-a-\(B2B\)-Salesman-104687.aspx](http://www.destinationcrm.com/Articles/Columns-Departments/Insight/Death-of-a-(B2B)-Salesman-104687.aspx).

[7] Gartner Research, "Magic Quadrant for Digital Commerce 2016 edition", 2016.

[8] M. Mam, E. Roegner, C. Zawada, *The Price Advantage*, New Jersey, Wiley & Sons, 2004.

[9] T. Poleto, V. D. H. de Carvalho, A. P. C. S. Costa, "The Roles of Big Data in the Decision-Support Process: An Empirical Investigation," *Proceedings of International Conference on Decision Support System Technology 2015 (Lecture Notes in Business Information Processing 216)*, 2015, pp. 10-21. <http://dx.doi.org/10.1007/978-3-319-18533-02>.

[10] N. Kowalczyk, P. Buxmann, "An ambidextrous perspective on business intelligence and analytics support in decision processes: Insights from a multiple case study," *Decision Support Systems*, vol. 80, 2015, pp. 1-13. <http://dx.doi.org/10.1016/j.dss.2015.08.010>.

[11] K. Kambatla, G. Kollias, V. Kumarc, A. Gramaa, "Trends in big data analytics," *J. Parallel Distrib. Comput.*, vol. 74, 2014, pp. 2561-2573. <http://dx.doi.org/10.1016/j.jpdc.2014.01.003>.

[12] S. Liu, A. H. B. Duffy, R. I. Whitfield, I. M. Boyle, "Integration of decision support systems to improve decision support performance," *Knowl Inf Syst.*, vol. 22, 2010, pp. 261-286. <http://dx.doi.org/10.1007/s10115-009-0192-4>.

[13] R. Phillips, *Why Are Prices Set The Way They Are?, Chapter 2, The Oxford Handbook of Pricing Management*, New York: Oxford University Press, 2014.

[14] A. Hinterhuber, S. M. Liozu, *Innovation in Pricing: Contemporary Theories and Best Practices*, 2nd ed., Abingdon-on-Thames, Routledge, 2017.

[15] M. Olszak, E. Ziemba, *Systemy Inteligencji biznesowej, jako przedmiot badań ekonomicznych*, ZN nr 113, Uniwersytet Ekonomiczny Katowice, 2012., pp. 13.

[16] B. Devlin, *Business UnIntelligence*, LLC, New Jersey, 2013.

[17] A. Elragal, "ERP and Big Data: The Inept Couple," *Procedia Technology*, vol. 16, 2014, pp. 242-249. <http://dx.doi.org/10.1016/j.protcy.2014.10.089>.

[18] H. Chen, R. H. L. Chiang, V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly*, vol. 36, no. 4, 2012, pp. 1165-1188.

[19] D. E. O'Leary, "Supporting decisions in real-time enterprises: automatic supply chain systems," *Inf Syst E-Bus Manage*, vol. 6, 2008, pp. 239-255. <http://dx.doi.org/10.1007/s10257-008-0086-0>.

[20] C.L. P. Chen, C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", *Information Sciences*, vol. 275, 2014, pp. 314-347. <http://dx.doi.org/10.1016/j.ins.2014.01.015>.

[21] H. Al-Aqrabi, L. Liu, R. Hill, N. Antonopoulos, "Cloud BI: Future of business intelligence in the Cloud," *Journal of Computer and System Sciences*, vol. 81, 2015, pp. 85-96. <http://dx.doi.org/10.1016/j.jcss.2014.06.013>.

[22] E. Gummesson, F. Polese, "B2B is not an island!", *Journal of Business & Industrial Marketing*, vol. 24, no. 5/6, 2009, pp. 337-350. <https://doi.org/10.1108/08858620910966228>.

[23] A. Smirnov, N. Shilov, A. Oroszi, M. Sinko, T. Krebs, "Product Knowledge Management Support for Customer-Oriented System Configuration," *Business Information System Workshops BIS 2017, Lecture Notes in Business Information Processing 303*, 2017, pp. 49-58. https://doi.org/10.1007/978-3-319-69023-0_5.

[24] A. Patel, B. Jaumard, "Design and Implementation of a Smart Quotation System," *Advances in Artificial Intelligence: Proceedings of 30th Canadian Conference on Artificial Intelligence, Canadian AI 2017, Edmonton, AB, Canada, Lecture Notes in Artificial Intelligence 10233*, 2017, pp. 191-202. https://doi.org/10.1007/978-3-319-57351-9_24.

[25] K. Senczyna, "The Use of Price Waterfall Model in Logistics", *Zeszyty Naukowe Politechniki Częstochowskiej Zarządzanie*, vol. 21, 2016, pp. 179-188.

[26] P. Kopal, D. Biswas, P. K. Chintagunta, J. Fan, K. Pauwels, B. T. Ratchford, J. A. Sills, "Retailer Pricing and Competitive Effects", *Journal of Retailing*, vol. 85, no. 1, 2009, pp. 56-70. <https://doi.org/10.1016/j.jretai.2008.11.005>.

[27] K. Sharma, I. Karlin, J. Keasler, J. R. McGraw, V. Sarkar, "Data Layout Optimization for Portable Performance", *Proceedings of Euro-Par 2015:*

Parallel Processing, LNCS 9233, 2015, pp. 250–262, https://doi.org/10.1007/978-3-662-48096-0_20

[28] S. van der Walt, S. C. Colbert, G. Varoquaux, “The NumPy Array: A Structure for Efficient Numerical Computation”, *Computing in Science & Engineering*, vol. 13, no. 2, 2011, pp. 22-30. <https://doi.org/10.1109/MCSE.2011.37>.

[29] W. Lieberman, “From yield management to price optimization: Lessons learned”, *Journal of Revenue and Pricing Management*, vol. 11, no. 1, 2011, pp. 40-43. <https://doi.org/10.1057/rpm.2010.44>.

[30] O. Roll, “Pricing trends from a management perspective”, *Journal of Revenue and Pricing Management*, vol. 8, no. 4, 2009, pp. 396-398. <https://doi.org/10.1057/rpm.2009.22>.

[31] C. Cizaire, “Pricing: The third business skill: Principles of price management”, *Journal of Revenue and Pricing Management*, vol. 13, no. 4, 2014, pp. 339-340. <https://doi.org/10.1057/rpm.2014.4>.

[32] T. L. Jacobs, R. Ratliff, B. C. Smith, “Understanding the relationship between price, revenue management controls and scheduled capacity – A price balance statistic for optimizing pricing strategies”, *Journal of Revenue and Pricing Management*, vol. 9, no. 4, 2010, pp. 356-373. <https://doi.org/10.1057/rpm.2010.18>.

[33] B. M. Noone, L. Canina, C. A. Enz, “Strategic price positioning for revenue management: The effects of relative price position and fluctuation

on performance”, *Journal of Revenue and Pricing Management*, vol. 12, no. 3, 2013, pp. 207-220. <https://doi.org/10.1057/rpm.2012.48>.

[34] A. E.-M. Bayoumi, M. Saleh, A. F. Atiya, H. A. Aziz, “Dynamic pricing for hotel revenue management using price multipliers”, *Journal of Revenue and Pricing Management*, vol. 12, no. 3, pp. 271-285. <https://doi.org/10.1057/rpm.2012.44>.

[35] A. A. Levis, L. G. Papageorgiou, “Active demand management for substitute products through price optimisation”, In: *Supply Chain Planning: Quantitative Decision Support and Advanced Planning Solutions*, Berlin: Springer, 2009. https://doi.org/10.1007/978-3-540-93775-3_4.

[36] D. Zhang, R. Kallelen, “Incorporating competitive price information into revenue management”, *Journal of Revenue and Pricing Management*, vol. 7, no. 1, 2008, pp. 17-26. <https://doi.org/10.1057/palgrave.rpm.5160120>.

[37] B.-N. Hwang, J. Tsai, H.-Ch. Yu, S.-Ch. Chang, “An effective pricing framework in a competitive industry: Management processes and implementation guidelines”, *Journal of Revenue and Pricing Management*, vol. 10, no. 3, 2011, pp. 231-243. <https://doi.org/10.1057/rpm.2009.47>.

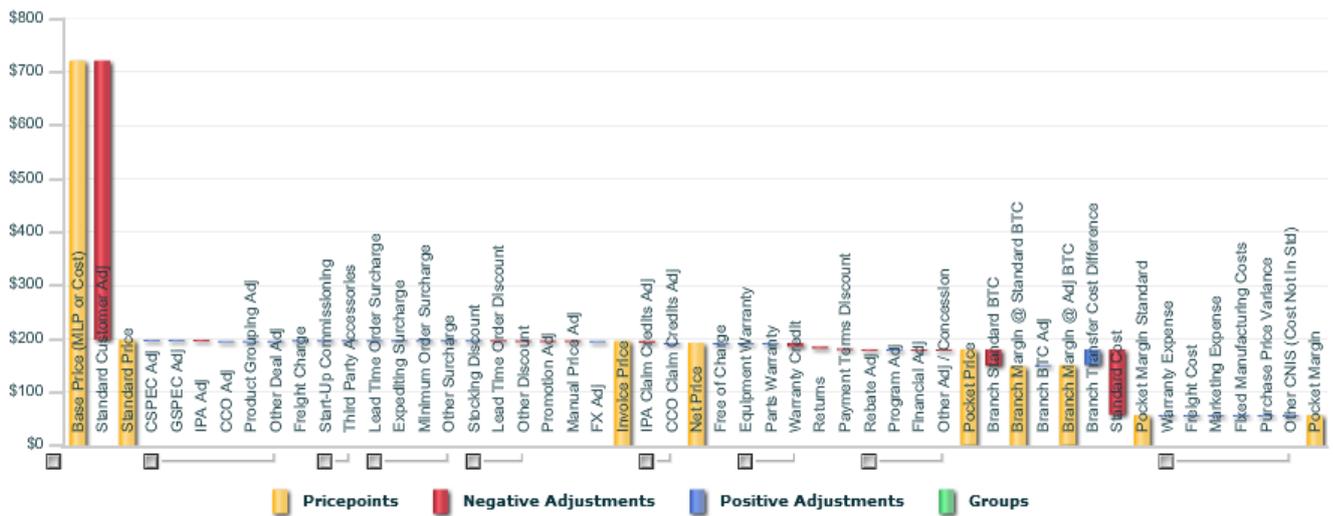


Fig. 5 A Price Waterfall generated within the CPQ solution from a sample dataset

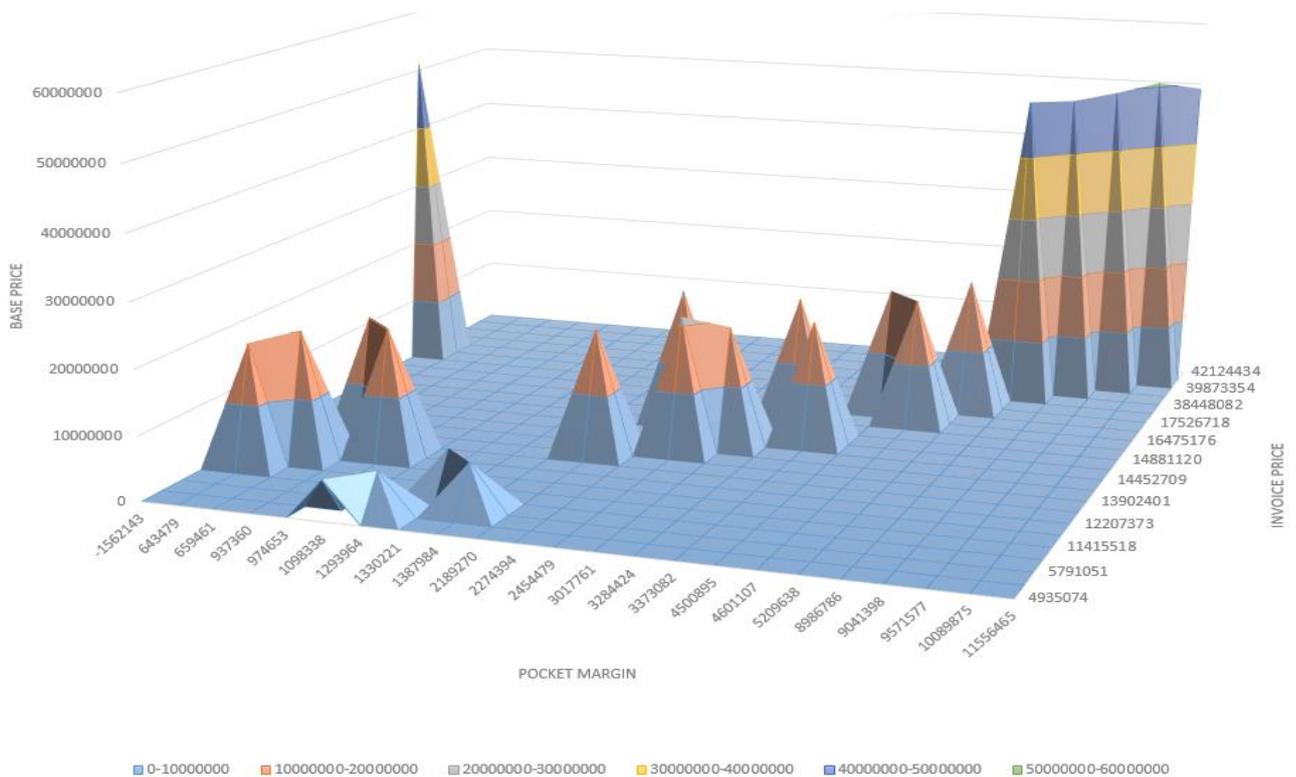


Fig. 6 3-D visualization of a sample Price Waterfall using sample dataset (presentation of a future functionality)

13th Conference on Information Systems Management

THIS event constitutes a forum for the exchange of ideas for practitioners and theorists working in the broad area of information systems management in organizations. The conference invites papers coming from two complimentary directions: management of information systems in an organization, and uses of information systems to empower managers. The conference is interested in all aspects of planning, organizing, resourcing, coordinating, controlling and leading the management function to ensure a smooth operation of information systems in an organization. Moreover, the papers that discuss the uses of information systems and information technology to automate or otherwise facilitate the management function are specifically welcome.

TOPICS

- Management of Information Systems in an Organization:
 - Modern IT project management methods
 - User-oriented project management methods
 - Business Process Management in project management
 - Managing global systems
 - Influence of Enterprise Architecture on management
 - Effectiveness of information systems
 - Efficiency of information systems
 - Security of information systems
 - Privacy consideration of information systems
 - Mobile digital platforms for information systems management
 - Cloud computing for information systems management
- Uses of Information Systems to Empower Managers
 - Achieving alignment of business and information technology
 - Assessing business value of information systems
 - Risk factors in information systems projects
 - IT governance
 - Sourcing, selecting and delivering information systems
 - Planning and organizing information systems
 - Staffing information systems
 - Coordinating information systems
 - Controlling and monitoring information systems
 - Formation of business policies for information systems
 - Portfolio management,
 - CIO and information systems management roles

- Information Systems for Sustainability
 - sustainable business models, financial sustainability, sustainable marketing
 - qualitative and quantitative approaches to digital sustainability
 - decision support methods for sustainable management

EVENT CHAIRS

- **Arogyaswami, Bernard**, Le Moyne University, USA
- **Chmielarz, Witold**, University of Warsaw, Poland
- **Karagiannis, Dimitris**, University of Vienna, Austria
- **Kisielnicki, Jerzy**, University of Warsaw, Poland
- **Ziemia, Ewa**, University of Economics in Katowice, Poland

PROGRAM COMMITTEE

- **Aguillar, Daniel**, Instituto de Pesquisas Tecnológicas de São Paulo, Brazil
- **Alghamdi, Saleh**, King Abdulaziz City for Science and Technology, Saudi Arabia
- **Bontchev, Boyan**, Sofia University St Kliment Ohridski, Bulgaria
- **Cingula, Domagoj**, Economic and Social Development Conference, Croatia
- **Czarnacka-Chrobot, Beata**, Warsaw School of Economics, Poland
- **Damasevicius, Robertas**, Kaunas University of Technology, Lithuania
- **Duan, Yanqing**, University of Bedfordshire, United Kingdom
- **El Emary, Ibrahim**, King Abdulaziz Univetrstity, Saudi Arabia
- **Espinosa, Susana de Juana**, University of Alicante, Spain
- **Feltus, Christophe**, Luxembourg Institute of Science and Technology, Luxembourg
- **Gawel, Aleksandra**, Poznan University of Economics and Business, Poland
- **Geri, Nitza**, The Open University of Israel, Israel
- **Halawi, Leila**, Embry-Riddle Aeronautical University, United States
- **Jankowski, Jaroslaw**, West Pomeranian University of Technology in Szczecin, Poland
- **Kania, Krzysztof**, University of Economics in Katowice, Poland

- **Kobyliński, Andrzej**, Warsaw School of Economics, Poland
- **Leyh, Christian**, Technische Universität Dresden, Germany
- **Michalik, Krzysztof**, University of Economics in Katowice, Poland
- **Mullins, Roisin**, University of Wales Trinity Saint David, United Kingdom
- **Muszyńska, Karolina**, University of Szczecin, Poland
- **Nuninger, Walter**, Polytech'Lille, Université de Lille, France
- **Ohira, Shigeki**, Nagoya University, Japan
- **Popescu, Elvira**, University of Craiova, Romania
- **Queirós, Ricardo**, Escola Superior de Media Artes e Design, Politécnico do Porto, Portugal
- **Rizun, Nina**, Alfred Nobel University, Dnipropetrovs'k, Ukraine
- **Rozevskis, Uldis**, University of Latvia, Latvia
- **Schroeder, Marcin Jan**, Akita International University, Japan
- **Sobczak, Andrzej**, Warsaw School of Economics, Poland
- **Swacha, Jakub**, University of Szczecin, Poland
- **Symeonidis, Symeon**, Democritus University of Thrace, Greece
- **Szczerbicki, Edward**, University of Newcastle, Australia
- **Travica, Bob**, University of Manitoba, Canada
- **Wątróbski, Jarosław**, University of Szczecin, Poland
- **Wielki, Janusz**, Opole University of Technology, Poland
- **Žemlička, Michal**, Charles University in Prague, Czech Republic

A new task scheduling approach based on Spacing Multi-Objective Genetic algorithm in cloud

Ali Belgacem
DCS Laboratory
EMP
bordj el bahri, Alger
al.BLgacem@gmail.com

Kadda Beghdad-Bey
DCS Laboratory
EMP
bordj el bahri, Alger
k.beghdadbey@gmail.com

Hassina Nacer
MOVEP Laboratory
USTHB
bab ezzouar, Alger
sino_nacer@yahoo.fr

Abstract—The dazzling progress in information and communication technologies, contributed significantly to the emergence of cloud computing paradigm, where it promotes prosperity in all fields of human activity, especially in business. Furthermore, manage the resources and use in ways that sharing with large number of users, consider as one of the challenges facing cloud computing environment today. Because cloud processes a huge tasks, which require the employment of scheduling techniques to handle and monitor the resources in an optimal, flexible and dynamic manner. In this paper, we review a new approach called Spacing-MOGA based on spacing distance to rank no-dominant solutions. It aims mainly to minimize both the makespan and cost of execution tasks on virtual machines (VMs). As well, we study its impact on the availability of resources. Experimental results show that S-MOGA is better than Max-min, PSO and MOGA methods, especially as it minimizes the number of active VMs.

Index Terms—Cloud computing. Resource allocation. Scheduling. Multi-Objective genetic algorithm. Spacing distance

I. INTRODUCTION

The emergence of Cloud computing is considered as a critical turning point in the world of computer, it made the computing power rentable. This announced the beginning of the fifth generation of computing after mainframe, personal computer, web and grid computing. In recent years, cloud has become very popular in different fields, especially for companies to increase economic efficiency and competitiveness. To meet every changing business, the companies need to invest time and budget to up their IT (Information Technology) infrastructure such as hardware, software and services. However, with on-premises IT infrastructure the scaling process is slow and the company is frequently unable to achieve efficient utilization of resources.

That is why, cloud computing is a paradigm shift that provide computing over Internet with an outstanding performance. It consists of set optimized virtual datacenters that provide various software, information and services (servers, storage, databases, networking, software, analytics and more). For use resources needed, companies can simply connected to cloud and use available resources on pay per use basis. This helps companies avoid capital expenditure on additional on-premises infrastructure resources and scale up or scale down according to business requirement.

Cloud computing environment is characterized by four types of accessing (public, private, hybrid and community), and offers three types of services (Software as a service (SaaS), Platform as a service (PaaS) and Infrastructure as a service (IaaS)). Besides, the virtualization technology is used to allocate the data center resources dynamically according to the application demands. Furthermore, live migration technology make it possible to assign each virtual machine to the physical machines while tasks are executing, which allow efficient utilization of resources. Other related technology that characterize cloud computing environment is the VM consolidation technique, it allows function many VMs on the same server in order to increase the number of unused servers.

On the other hand, resources are an entities where tasks are allocated. Each resource has its own characteristics (computing power (CPU), memory size, etc.). In addition, there are different types of resources: storage resources, power resources, networking resources and compute resources. Since, the scheduling mechanism is an important issue that improves the use of resources and also makes a better performance of this computing environment, many approaches are used to find optimal solutions and to achieve the preferable results.

Most of the studies about the techniques used in the area of task scheduling in cloud, focused on the application of heuristic and meta-heuristic mechanisms [1]. They are a nature inspired algorithm based on the biological or physique phenomena. For example, the work [8] discussed the techniques used for task scheduling founded on Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Genetic Algorithm (GA), and two novels League Championship Algorithm (LCA) and BAT algorithm. In addition to this, the authors in [9] illustrated an analysis about the workflow scheduling approaches based Simulated Annealing (SA), and Cat Swarm Optimization (CSO).

The remainder of the paper is organized as follows: section II illustrates the existing approaches of task scheduling problems in cloud. The description of proposed algorithm is given in section III. Section IV presents the formulation of the studied problem. The proposed algorithm steps are explained in section V. The section VI describes the simulation methodology used to evaluate our approach. Finally section VII presents the conclusion of the paper.

II. RELATED WORKS

There are several approaches applied to examine the task scheduling in cloud, with a view to solve various resources allocation problems [1]. Bey, K.B., et al. presented a new scheduling strategy based on load balancing (LBE) approach for an independent tasks which gave a good results in terms of execution time, makespan and resource utilization [2]. The reference [4] introduced a pareto-based multi-objective workflow scheduling algorithm for allocating different on-demand instances with optimal makespan and various prices. But, authors didn't consider the security issues, reliability of spot instances or energy consumption of the system. Correspondingly, the paper [7] aimed to minimize the makespan and ensure a better load balancing of system. The proposed approach assigns the tasks without deadlines or priorities which provided better results than NSGA-II algorithm.

In the work [11] Portaluri, Giuseppe et al. applied MOGA algorithm to reduce the power consumption in data center. The authors evaluate the proposed algorithm by combining between different numbers of auxiliary objectives which affected negatively on quality of results. Moreover, they did not consider the traffic exchanged between VMs. Instead, Zhang, Fan et al. [12] presented a multi-objective scheduling (MOS) scheme for the multitasking workflow application over different virtual clusters. It allows mainly to reduce scheduling overhead time and yet a close to optimal performance. However, this method applied just for a small number of nodes.

In order to optimize both makespan and cost Zhu, Zhaomeng et al. proposed Evolutionary Multi-objective Optimization (EMO)-based algorithm that contain novel encoding scheme to represent the genetic operators. This algorithm gave more stability on the workflow scheduling problem, without conceder more than one pricing schemes, instance type groups or even multi-clouds in a single schedule [13]. Other works based on ant colony algorithm, as the researche [14], the authors take into account the makespan, cost, deadline violation rate, and resource utilization as constraints to achieve a multi-objective optimization of both makespan and cost. The applied method is better than other similar methods results in terms of makespan.

Our major contributions in this work is to impose an integrated solution that covering several aspects of resource allocation in cloud, as following:

- minimize makespan
- minimize cost

III. PROBLEM DESCRIPTION AND DEFINITIONS

Before solving our studied problem, it is important to define the main actors and the architecture of the task scheduling system in cloud, as shown in the following:

A. Definition

This problem considers the following difinitions:

Task (T): It reflects a set of independent requests $T = \{t_1, t_2, \dots, t_m\}$ that describe the client's requirement, each task is characterised by identifier id_t and *resource requirment*

$(T_{FileSize_i}, T_{CPU_i})$.

Virtual Machine (VM): Is a VM image hosted on cloud infrastructure (exactly on servers). It may contain an OS, data files, and applications. Each VM instance is represented by its identifier id_v and *resources available* (V_{M_j}, V_{CPU_j}) .

Server Manager: It provides a centralized platform for managing the set of VMs in data center, allowing to create and deploy a VMs on physical servers quickly and easily. Also, it contains the scheduling mechanisms.

Host: It is a server for hosting the VMs.

B. The System Model

As shown in figure 1, users send the requests to cloud provider to express their needs. The server manager analyse the request in order to extract the resources requirement, then assigns the request to the available resources. So that, the provider must serve the customer in an optimal way that meets his requirements. In other words, the main steps of the stadied problem are:

- The client transmits a request to determine its requirements for resources, via a user interface (figure 1).
- Examining and revising the client request for analysing and evaluating how the required service could be provided at a lower cost and in short time (occurs at the level of Cloud Broker).
- The scheduler assigns the tasks to apropiet VMs.

This study deals with the mapping between a set of independent tasks and set of VMs. The VMs are hosted on physical machines (PM). So that, the problem is modeled as follows:

Input:

- Set of $VM_s = \{V_{m1}, V_{m2}, \dots, V_{mn}\}$ with different configurations such as CPU type and memory size. Each machine represented by their (id) and millions of instructions per second (mips).
- Set of independent tasks $T = \{t_1, t_2, \dots, t_m\}$ with different sizes.

Output: The best mapping of T_i to VM_j (T_i, VM_j), in manner to reduce both makespan and cost .

IV. PROBLEM FORMULATION

A. The resource cost model

Cost: Our concern here is the cost of resources reservation. In the problem addressed, the cost is expressed as follows:

$$C_i(t_{run_i}) = c_j \times t_{run_i} + C_{tr_{ij}} \quad (1)$$

$$C_{tr_{ij}} = \frac{T_{FileSize_i}}{mips_j \times 32} \times \varepsilon, \quad \varepsilon = 0.001\$ \quad (2)$$

$$Total \ cost = \sum_{i=0}^m C_i \quad (3)$$

Where C_i is the cost associated with the execution of task i on the resource j , c_j is the price of using resource j , t_{run_i} is the duration time of running the task i on resource j , $C_{tr_{ij}}$ is the transfer rate cost for a bus of 32 bits wide with a clock

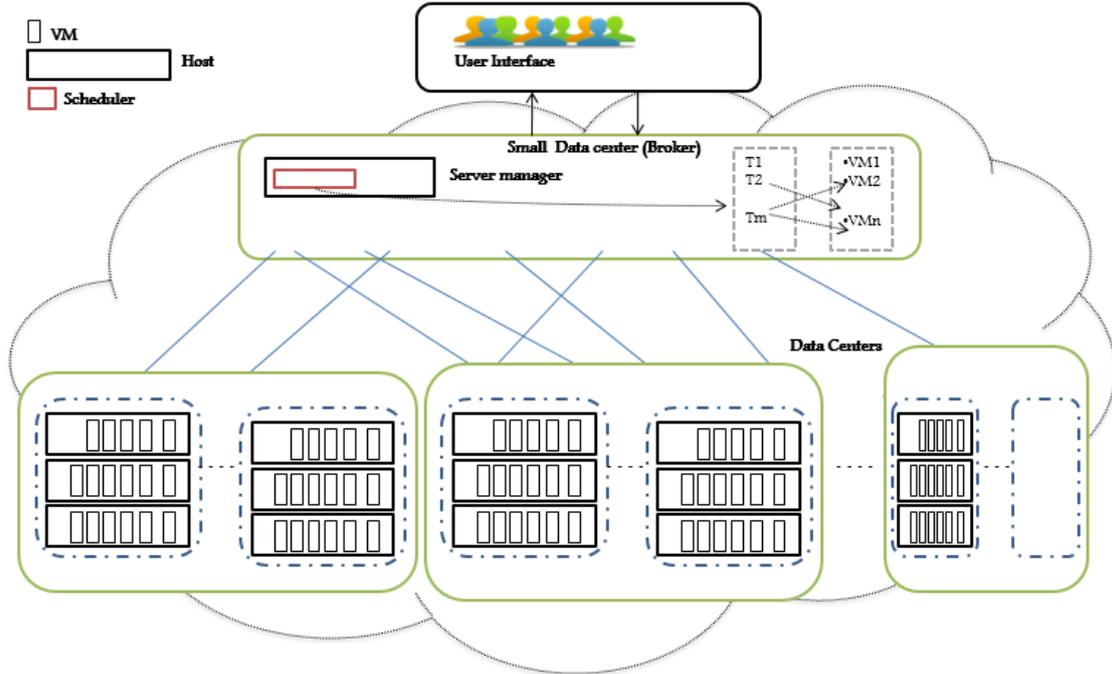


Fig. 1. System architecture

speed of $mips_j$, ε is the transfer price and m is the number of all tasks.

B. The Computation time (makespan) model

Makespan: It is the total execution time of all the tasks. In this work, the expression of makespan is given as follows:

$$ET_{ij} = \frac{T_{length_i}}{mips_j} \tag{4}$$

$$Makespan = \sum_{i=0}^m ET_i \tag{5}$$

Where ET_{ij} is the execution time of task i on resource j .

C. The function objectives

$$Minimize \theta(x) = \varphi(x), \phi(x) \tag{6}$$

Subject to :

$$V_{Mj} \geq T_{FileSize_i} \tag{7}$$

$$V_{CPUj} \geq T_{CPU_i} \tag{8}$$

$$x \geq 0 \tag{9}$$

Here x is a feasible solution (execution of task i on VM instance j), $\varphi(x)$ is a function of the performance objectives that refer to makespan. $\phi(x)$ is the objective function of the user budget costs. The equation 7 and 8 means that the CPU and memory configuration of VMs must be greater than or equal to the user request requirement.

V. PROPOSED ALGORITHM

In order to review our multi-objective method, firstly the fundamentals of the proposed algorithm should be discussing. Then explaining the S-MOGA algorithm steps, as follows:

1) *Pareto Dominance:* When solving a problem of multi-objective optimization, a multitude of solutions be obtained. Only a limited number of these solutions will interest us. For a solution to be interesting, there must be a relation of dominance between the considered solution and the other solutions. More precisely, the vector $\vec{x} = (x_1, x_2, \dots, x_n)$ dominates the vector $\vec{y} = (y_1, y_2, \dots, y_n)$ if:

- \vec{x} is at least as good as \vec{y} in all objectives,
- \vec{x} is strictly better than \vec{y} in at least one objective.

The solutions that dominate, but do not dominate each other are called no-dominated solutions.

2) *Genetic Algorithm:* It is a search algorithm based on directed random searches to locate optimal solutions. It is meta-heuristic based on the iterative application of stochastic operators on a population of candidate solutions [10].

3) *Multiple Objective Genetic Algorithm (M.O.G.A):* This method is based on pareto dominance. The "rank" of an individual (order number which ranks an individual in relation to others) is given by the number of individuals who dominate it, in each iteration.

4) *S-MOGA:* Our proposed approach based on MOGA, but it applies a new way of individual ranking based on the Spacing Distance [5]. The Spacing-MOGA process explained in the following steps :

a) *Stop criteria:* To determine the stop criteria, Ω defined as stabilizing factor, it increases when the value of maksepan

in the current iteration is the same in precedent iteration, and it take $\Omega = 0$ in otherwise.

b) *Initialisation*: The first step is beginning with a set of individuals which is called a population. Each individual is a solution to the problem. In the studied problem the genes consider as tasks and the positions of a genes as positions of VMs. This phase aims to dispatch the selected task to a randomly selected available VMs.

c) *Crossover*: In this phase of our proposed algorithm, for each pair of chromosome to be mated, a crossover point is chosen at middle of chromosome from within the genes.

d) *Mutation*: In new offspring formed after crossover phase, some of the individuals be flipped randomly in the population string.

e) *Mixed population*: The initiale population and sorte population are mixed, to form a *mixed population* after the generation of new solutions. Then, the individuals in the mixed population are hierarchically categorised into the dominant and no dominant subsets based on the concepts of dominance mentioned above.

f) *Rank*: The method inspired by the Spacing Distance used to determine the relationship between two no-dominance solutions, where the following formula is applied:

- Calculate the ω_s (the average distance for each φ_s):

$$\omega_s^{nd} = \frac{\varphi_s^{max} - \varphi_s^{min}}{\sigma - 1} \quad (10)$$

where ω_s^{nd} is the average distance for the no-dominat set, σ is the number of non dominant individuals. The following equation shows how calculate the distance per individual from its neighbors:

$$d^{nei}(i) = \frac{|\varphi^i - \varphi^{i-1}| + |\varphi^{i+1} - \varphi^i|}{2} \quad (11)$$

- Measure the new fitness of each individual:

$$\varphi(i) = d^{nei}(i) - \omega_s^{nd} \quad (12)$$

g) *Filter Solutions*: This step updates our population set, through changing just the solutions that have a worst fitness in the enhanced population pool compared to those achieved by the above steps, in the same iteration.

Figure 2 shows the Organogram of S-MOGA algorithm. The pseudo code of S-MOGA is presented as algorithm 1.

VI. IMPLEMENTATION

A. Simulation Environment

The platform used to execute the experiments was an Intel I5 3320M 2.60 GHz equipped with 4GB RAM with OS Windows 7 Professional. The experiments programmed with Cloudsim toolkit in Eclipse development environment, for modeling and simulation of cloud computing infrastructures and resource allocation [3]. It is a simulator tool founded on Java application, which is an object-oriented computer programming language and a portable computing execution environment [6].

Algorithm 1 S-MOGA

Input : $T_1, T_2, \dots, T_m, VM_1, VM_2, \dots, VM_n$

Result: An optimized generated schedule

$\Omega := 0$;

while $\Omega \leq 10$ **do**

function INITIALISATION()

function CROSSOVER()

function MUTATION()

function MIXED POPULATION()

function RANK ()

function FILTER SOLUTIONS ()

end

Return the best mapping (Tasks, VMs) as the best found solutions.

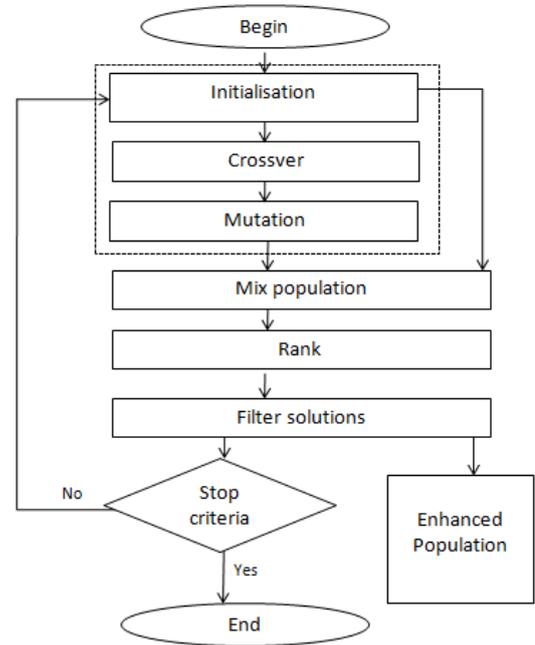


Fig. 2. Organogram of the S-MOGA

B. Experimentation Results

The experiments, mainly focus to evaluate the makespan of our proposed algorithm. Also, study the variation of budget costs for resource utilization. Besides, compare the proposed algorithm of this paper with the original Particle Swarm Optimization (PSO) algorithm, the classical heuristic algorithm Max-Min and MOGA scheduling.

The test is done in two cases: in the first case assumes that the number of VMs is 16789. In the second case considers that the number of VMs is 34568. The MIPS of each VM is between [2000, 2050]. The length of tasks is between [100, 1070]. The configuration of image size, VM memory and VM bandwidth is illustrated in Table I. Furthermore, the number of tasks varying as 1000, 1678, 2874 and 3456. In addition, the test considers the reservation cost of various VMs 0.02 \$/hour.

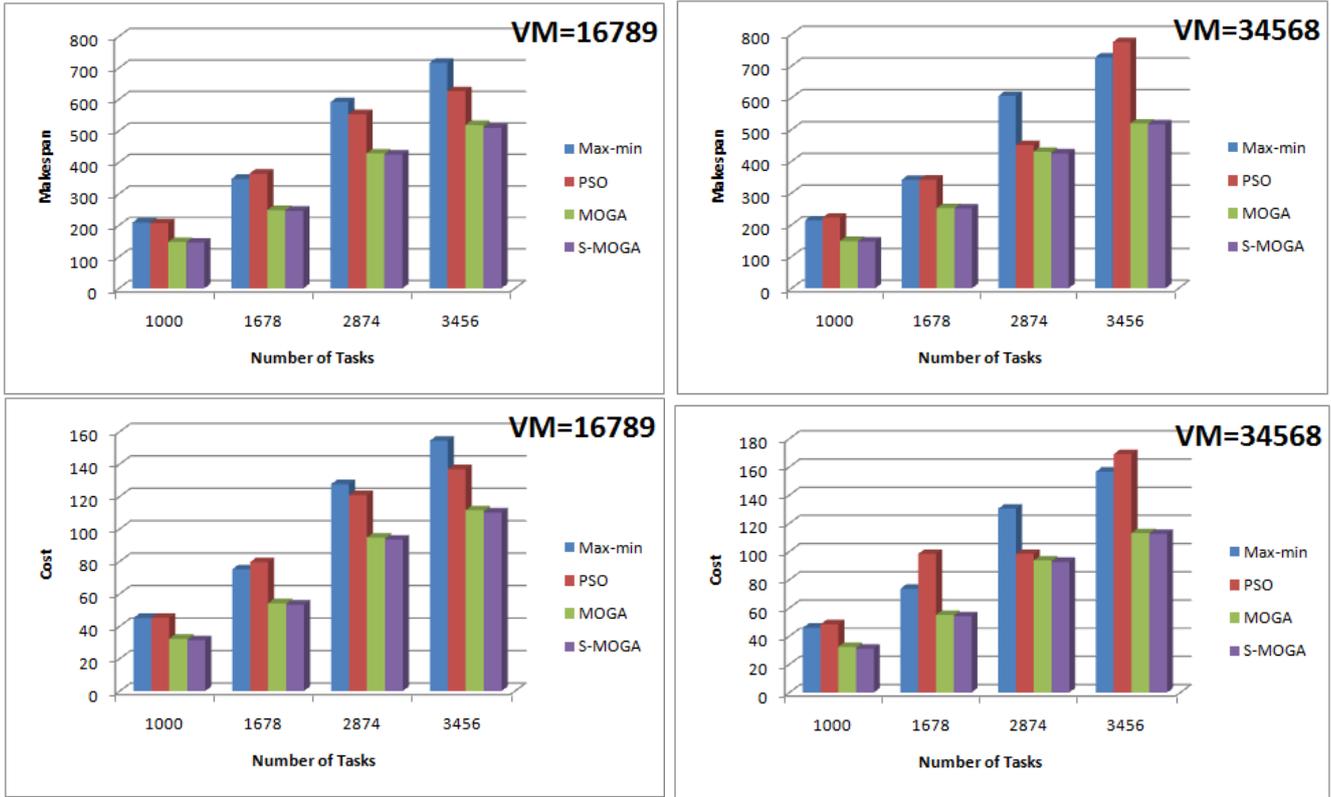


Fig. 3. Comparisons of proposed scheduling algorithm with Max-min, PSO and MOGA algorithm for Makespan and Cost

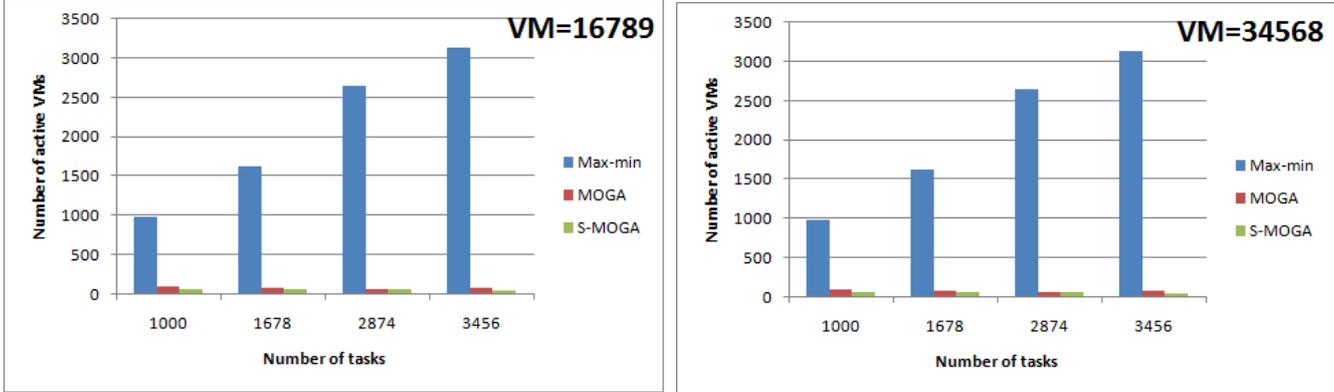


Fig. 4. Comparisons of proposed scheduling algorithm with Max-min and MOGA algorithm for Availability of Resources

TABLE I
THE PARAMETER SETUP OF VMS

size	ram	mips	bw
10000 MB	[500, 512]	[2000, 2050]	1000

In order to evaluate the availability of resources aspect of our proposed algorithm, the next test keeps the same range of MIPS and task length variation. Also, defines the Availability

of Resources (AR) of the system as a number of *inactive VM* in each allocation. Hence, considers that the AR is increased only if the number of the VM active in the system is minimized, which means a good availability of the resources for next allocation. So that, the AR of the system is calculated as shown in the following equation:

$$AR = m - \sum_{j=0}^k Av_j \tag{13}$$

Where Av_j is an active VM_j and k is the total number of

TABLE II
SIMULATION RESULTS OF MAKESPAN AND COST FOR MAX-MIN, PSO, MOGA AND PROPOSED ALGORITHM

Tasks	Lengthe	VM	MIPS	Max-min		PSO		MOGA		S-MOGA	
				Makespan	Cost	Makespan	Cost	Makespan	Cost	Makespan	Cost
1000	[200,300]	9087	[2340,3450]	89.130	20.247	82.11	18.968	61.142	14.126	60.699	14.018
2345				215.897	48.931	191.323	44.283	142.760	32.983	140.558	32.548
3456				316.229	71.758	277.063	64.028	210.747	48.707	208.346	48.188
4000				369.269	83.677	306.841	70.8797	244.300	56.4797	241.308	55.789
1000	[900,2345]	10056	[12340, 45670]	71.815	14.689	58.477	12.002	35.798	7.3503	35.706	7.325
2345				165.3199	33.805	131.782	27.055	84.677	17.3792	83.792	17.200
3456				246.355	50.373	178.085	36.551	123.990	25.453	122.844	25.217
4000				281.602	57.597	190.499	39.096	142.753	29.308	142.723	29.295
1000	[9000,22556]	30974	[56720, 345960]	116.048	23.264	86.391	17.326	50.760	10.1795	50.9299	10.213
2345				268.527	53.829	168.676	33.828	119.206	23.905	116.483	23.3598
3456				403.235	80.833	356.743	71.5409	174.846	35.064	173.772	34.848
4000				464.007	93.017	305.234	61.211	202.784	40.666	200.098	40.129

TABLE III
SIMULATION RESULTS OF AVAILABILITY OF RESOURCES FOR MAX-MIN, MOGA AND PROPOSED ALGORITHM

Tasks	Lengthe	VM	MIPS	Max-min	MOGA	S-MOGA
				VM active	VM active	VM active
2354	[100, 500]	1000	[234, 567]	896	27	4
3785				972	34	8
4000				983	38	6
2354	[1950, 2700]	3789	[7567, 8464]	1747	2829	20
3785				2403	2818	17
4000				2450	2819	19

Av, m is the total number of VMs.

1) *Makespan and Cost*: As it is shown in figure 3, the S-MOGA approach has a lower values of makespan and cost compared to the other approaches Max-min, PSO and MOGA.

The experiment is repeated with varying: the MIPS value of VMs at different intervals, for various numbers of tasks and VMs. So that, the simulation takes 1000, 2345, 3456 and 4000 numbers of tasks run on 9087, 10056 and 30974 heterogeneous machines successively in a cloud (Table II).

The results illustrated in the table II indicate clearly that the proposed scheduling algorithm performs well compared with other algorithms, it gave significant improvements in terms of

makespan and cost of resource reservation.

2) *Availability of resources*: To examine the impact of our proposed algorithm on the AR, the experiment is reoccurred by using a new configuration of VMs: different intervals of mips, various numbers of tasks and VMs. The test done for 1000 and 3789 heterogeneous VMs successively (Table III).

The figures 4 and the results of table III show that the proposed algorithm allows using a minimum of resources by consolidate the execution of tasks over few number of VMs and make off others which ensure the availability of resources.

VII. CONCLUSION

This paper presents an improvement of Multi objectives Genetic Algorithm based on spacing distance (S-MOGA) to enhance the Quality of Service (QoS) requirements in Cloud by minimizing the total task execution time and cost. We considered various independent tasks with different lengths, which are corresponding to user request and various VMs to mimic resource allocation in cloud. The experimental results shown that the proposed algorithm generated results better to those of the MOGA, PSO and Max-min in terms of makespan and cost. In addition, we studied the performance of the Spacing-MOGA from the viewpoint of their feasibility to meet the needs of users. Where the results confirmed that the proposed algorithm offering a good availability of resources compared with other methods.

In future work, we will consider other scheduling algorithm that can be used to solve a resource allocation problem in a Cloud computing environment. As well, we will study its ability to manage the resources in a dynamic manner.

REFERENCES

- [1] Ali Belgacem, Kadda Beghdad-Bey, and Hassina Nacer. Task scheduling in cloud computing environment: A comprehensive analysis. In *International Conference on Computer Science and its Applications*, pages 14–26, Algiers, Algeria, 24-25 April 2018. Springer.
- [2] Kadda Beghdad Bey, Farid Benhammadi, Mohamed El Yazid Boudaren, and Salim Khamadja. Load balancing heuristic for tasks scheduling in cloud environment. In *Proceedings of the 19th International Conference on Enterprise Information Systems – Volume 1: ICEIS.*, pages 489–495, April 26-29, in Porto, Portugal, 2017. INSTICC, SciTePress.
- [3] Rodrigo N Calheiros, Rajiv Ranjan, Anton Beloglazov, César AF De Rose, and Rajkumar Buyya. Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and experience*, 41(1):23–50, 2011.
- [4] Juan J Durillo and Radu Prodan. Multi-objective workflow scheduling in amazon ec2. *Cluster computing*, 17(2):169–189, 2014.
- [5] L Falahiazar and H Shah-Hosseini. Optimisation of engineering system using a novel search algorithm: the spacing multi-objective genetic algorithm. *Connection Science*, pages 1–17, 2018.
- [6] Tarun Goyal, Ajit Singh, and Aakanksha Agrawal. Cloudsim: simulator for cloud computing infrastructure and modeling. *Procedia Engineering*, 38:3566–3572, 2012.
- [7] Ashish Gupta and Ritu Garg. Load balancing based task scheduling with aco in cloud computing. In *Computer and Applications (ICCA), 2017 International Conference on*, pages 174–179, Doha, United Arab Emirates, 6-7 Sept 2017. IEEE.
- [8] Mala Kalra and Sarbjeet Singh. A review of metaheuristic scheduling techniques in cloud computing. *Egyptian informatics journal*, 16(3):275–295, 2015.
- [9] Mohammad Masdari, Sima ValiKardan, Zahra Shahi, and Sonay Imani Azar. Towards workflow scheduling in cloud computing: a comprehensive analysis. *Journal of Network and Computer Applications*, 66:64–82, 2016.
- [10] Mohand Mezmaz, Nouredine Melab, Yacine Kessaci, Young Choon Lee, E-G Talbi, Albert Y Zomaya, and Daniel Tuytens. A parallel biobjective hybrid metaheuristic for energy-aware scheduling for cloud computing systems. *Journal of Parallel and Distributed Computing*, 71(11):1497–1508, 2011.
- [11] Giuseppe Portaluri and Stefano Giordano. Multi objective virtual machine allocation in cloud data centers. In *Cloud Networking (Cloudnet), 2016 5th IEEE International Conference on*, pages 107–112, Pisa, Italy, 3-5 Oct 2016. IEEE.
- [12] Fan Zhang, Junwei Cao, Keqin Li, Samee U Khan, and Kai Hwang. Multi-objective scheduling of many tasks in cloud platforms. *Future Generation Computer Systems*, 37:309–320, 2014.
- [13] Zhaomeng Zhu, Gongxuan Zhang, Miqing Li, and Xiaohui Liu. Evolutionary multi-objective workflow scheduling in cloud. *IEEE Transactions on parallel and distributed Systems*, 27(5):1344–1357, 2016.
- [14] Liyun Zuo, LEI Shu, Shoubin Dong, Chunsheng Zhu, and Takahiro Hara. A multi-objective optimization scheduling method based on the ant colony algorithm in cloud computing. *IEEE Access*, 3:2687–2699, 2015.

Data quality evaluation: a comparative analysis of company registers' open data in four European countries

Janis Bicevskis
Faculty of Computing
University of Latvia
Latvia
Janis.Bicevskis@lu.lv

Zane Bicevska
DIVI Grupa Ltd
Latvia
Zane.Bicevska@di.lv

Anastasija Nikiforova
Faculty of Computing
University of Latvia
Latvia
AN11093@lu.lv

Ivo Oditis
Faculty of Computing
University of Latvia
Latvia
Ivo.Oditis@lu.lv

Abstract—This paper is devoted to the analysis of open data quality of the company registers in four different countries. The data quality evaluation was obtained using a methodology that involves the creation of three-part data quality model: (1) the definition of a data object to analyse its quality, (2) data object quality specification using DSL, (3) the implementation of an executable data quality model enabling the scanning of a data object and detecting its deficiencies. All three components of the data quality model are designed as graphical language families, which allow formulating data quality specification for non-IT professionals. Validation of an open data published by company registers in four different European countries shows deficiencies in the published data and demonstrates the applicability of the proposed methodology for data quality evaluation.

Index Terms—data quality, executable models, Company register

I. INTRODUCTION

THE open-world model is gaining ever more popularity [1]. The society calls for direct access to the information avoiding mediators and filters. The process of making data freely available to the public includes also providing access to the data in public company registers that serve public administration purposes. The state institutions, business and individuals are interested in facilitating the process of effective communication and receiving services. This is not possible without the accurate and timely registration of objects such as population, real estate, vehicles, taxes and other objects legally required to register in public registers.

In the situations when state information system data is made public, data quality is of crucial importance, i.e., can the open data be trusted and used, what are recommended purposes of data usage.

Guidelines and key principles for development of state information system in Latvia were defined in national program "Informatics" more than 17 years ago [2]:

- Public objects shall be accounted for and registered by a public agency operating under the supervision of the relevant ministry. The agency shall be responsible for the registered data including data precision, completeness, timeliness, etc. Data on each public object shall be registered by only one public agency.

- Duplicate entry or the same public object in another public agency is prohibited. All matters necessary shall be administered by the agency where the data object is registered.
- Information on public object shall be recorded at the time when information is generated avoiding temporary recording on paper and later data entry into an information system.
- Documents certifying public object registration shall be printed from the information system data base, thus ensuring compliance between the printed documents and the data stored in the information system.

Many state information systems were developed and implemented according to these principles that were meant to provide accumulation of qualitative information into public registers.

Assessing the state information systems in Latvia, we have to acknowledge that, indeed, many information systems operate in accordance with these principles. For example, personal identification documents are printed from the population register, vehicle registration certificates and driver's licenses are printed from the vehicle and driver register while using personal data from the population register, etc., Unfortunately, not all state information systems have implemented these principles, therefore it is important to analyze the data quality of these systems. Since state information systems have restricted accessibility, the task for researchers was to create an independent "external" mechanism for assessing the data quality without using the same information system that accumulated data.

Numerous studies have led to various definitions of data quality. For instance, data are of good quality if they satisfy the requirements imposed by the intended use [3]. The ISO 9001:2015 standard [4] considers data quality as a relative concept, largely dependent on specific requirements resulting from the data use. The same data may be sufficiently qualitative in one situation but completely useless under other circumstances.

This principle is confirmed by analyzing the data quality of the Latvian Population Register in 1999. In the Latvian Population Register a person is described by 7 data groups. Such personal identification data as registration number, name,

and surname contained a relatively small number of errors and were assessed as qualitative. While the data on place of residence, links to parental data or links to personal data of children were far from desirable quality. The data on place of residence available in the Latvian Population Register could not be used to contact the persons. Moreover, due to insufficient infrastructure and internet availability, the data entry into the register was not timely.

Due to the fact that the Population Register contains not freely published personal data, the study focused on an analysis of publicly available data of the company registers in 4 countries (Latvia, Estonia, Norway and the United Kingdom). Company records with the values of the parameters were "scanned", and deviations from data quality specifications were registered. The produced results are quite surprising showing that data accumulated and published for many years is of dubious quality.

This paper has two main objectives. The first objective is to clarify whether publicly available data provided by company register is trustful and what is the quality of these data for simple use, for instance, identifying a company and sending a message to this company. Our second objective is to verify quality evaluation methodology with the help of executable data quality models described further in detail [5].

The paper deals with following issues: overview about the methodology of evaluation of data quality (Section 2), an analysis of data quality of company registers in four countries – Latvia, Estonia, Norway, and the United Kingdom (Section 3).

II. METHODOLOGY

So far, many studies have been devoted to data quality evaluation methodology and practice. All these studies can be divided into several groups:

- General studies on the data quality, in most cases, defining the data quality dimensions and their groupings [3], [6], as well as evaluation methodologies [7], [8]. The sources [6] and [8] provide a comprehensive overview on existing researches, methodologies and tools which can be explored and/or used in other researches on data quality problem.
- The specific industry related data quality evaluation by analyzing industry-specific data and evaluation methodology [7], [9], [10]. Methodologies mentioned in [7], [10] are insufficiently industry-specific, as a result, it is difficult to apply their results to another area (it is difficult to re-use them for customizing to specific use-cases). Most of the proposed guidelines in the data quality evaluation methodology defined in [9] are difficult to use (especially for non-IT specialists) as it requires a lot of resources to complete required specification tables, for example, data quality parameter specification. Moreover, it could require involvement of the authors of this methodology. Therefore, it can hardly be used widely despite the fact that using this methodology could ensure comprehensive data quality specification and efficient data quality analysis.

- The data quality evaluation of partially structured (semi-structured) data [11] and poorly structured data, such as Wikipedia [12]. Methodology for data quality analysis proposed in [11] covers even data quality improvement phase and allows involving stakeholders. Unfortunately, this methodology is not easy to understand for stakeholders. Moreover, the stakeholders are involved considering only their needs (finding out and satisfying them) and not trying to involve them in specific stages of the process of data quality analysis. Another method for quality evaluation is described in [12] and used to evaluate quality of Wikipedia articles and information contained in their info boxes; however, it is not applicable to specific datasets (e.g. data storing in the relational databases) without converting them into the appropriate format.

The quality of open data published by the company registers in four European countries will be evaluated using the approach described in [5], which is used to evaluate the quality of fully-structured data. This chapter gives an insight into the methodology of data quality models according to the approach described in [5]. It is characterized by the following main characteristics:

- For each specific application, you can create your own data quality model and evaluate the quality of data for a particular application.
- Data quality model can be described at different abstraction levels from informal text in natural language to an automatically executable model, SQL statements or program code.
- Data quality model consists of three types of graphical charts describing data objects, data quality specification and data quality evaluation processes. The charts could be configured by creating and using domain-specific languages.
- The data quality model is "external" to the information system that stores the accumulated data, i.e., the data quality model can be built without knowing about technologies used for accumulation of data.

The proposed methodology is useful for both developers of information systems for defining data quality specification and for industry experts to assess the quality of the data published by different data suppliers.

The proposed data quality evaluation solution consists of 3 main components: (1) data object, (2) quality requirements, and (3) quality evaluation process. These components form data quality model. The data object description defines the data which quality must be analyzed, the quality specification defines conditions which must be met to admit data as qualitative, and the description of quality evaluation process defines the procedure that must be performed to evaluate data quality.

A. Data Object

Traditionally the notion of a data object is understood as the set of values of the parameters that characterize a real-life object. The research results will be illustrated by simple examples from the Company Register of Latvia, the quality of which will be analyzed in the next chapter. In Fig.1 the data object Company is depicted with its attributes: Reg_number – registration number of company, Name – name of company, Type – type of company, etc. The description of company is partly formal as rules for attribute values syntax are given.

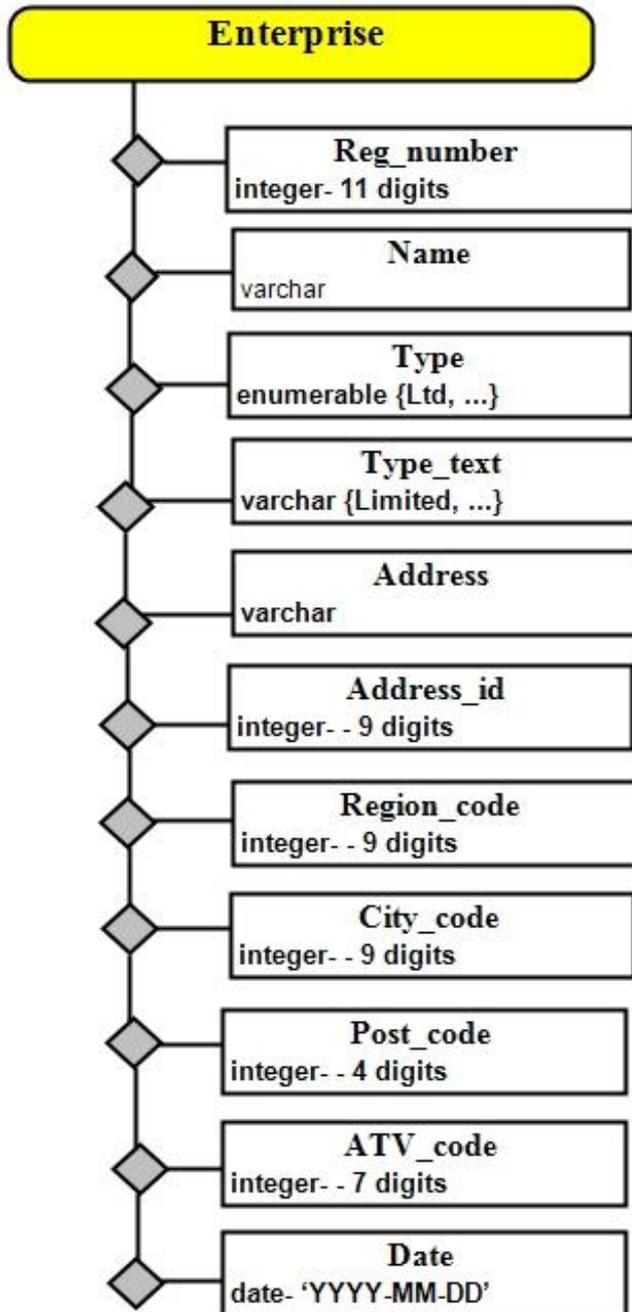


Fig.1. Data object “Company”

The quality checking of any data object parameters’ value is reduced to an examination of individual values’ properties, for instance, checking whether a text string may serve as a value of the field Name, or whether a value of the field Address is a correct address. Anyway, the checking of parameter values is a local and formal process. At the current stage of the research, it does not respect contextual interlinks with other data objects and does not check the compliance of data with the true characteristics of a real company.

The syntax rules for the permissible values of the data object’s fields can be formulated at different levels of abstraction including the formal language grammar and definitions of variables in programming languages. In the latter case, the data object model is closely related to the environment in which the model will be implemented.

A specific quality control of a particular data object usually is a part of the input data quality control in every information system. Data is usually entered into an information system by filling in screen form fields, followed by an information quality check and its retention in the database. In cases when the input fields are not filled correctly, the user receives an error message and may adjust the input data.

Information systems deal not only with individual data objects but also process many data objects in a unified way. In this case, the classes of data objects are used they represent many objects of the same structure. A data object class has a name, and its elements have the same structure and the same characteristic parameters. Each individual data object may contain parameter values fully or partially.

B. Data Object Quality Specification

A data quality specification contains conditions that must be met in order a data object is considered of high quality. The quality specification (Fig. 2) may contain informal descriptions of conditions, for example, in natural language or formalized implementation-independent descriptions. Data quality specification of a data object is defined by logical expressions. The names of data object’s attributes/ fields serve as operands in the logical expressions. The traditional means of programming languages, for example, the programming language C #, may be used as operations.

When processing the data object class, data object class instances are selected from the sources of information and written into a collection. All instances were processed cyclically by examining the fulfilment of a quality specification of each individual instance. The quality specification was similar to the specification used in the processing an individual data object. Thus, the quality problems of each individual instance were identified.

C. Quality evaluation process

The first step in the quality evaluation process describes the activities to be taken to select data object values from data sources. Thereafter, one or more steps are taken to evaluate the data object with a specific quality of the data, each of which describes one test for the compliance of data object “Company” with the quality specification.

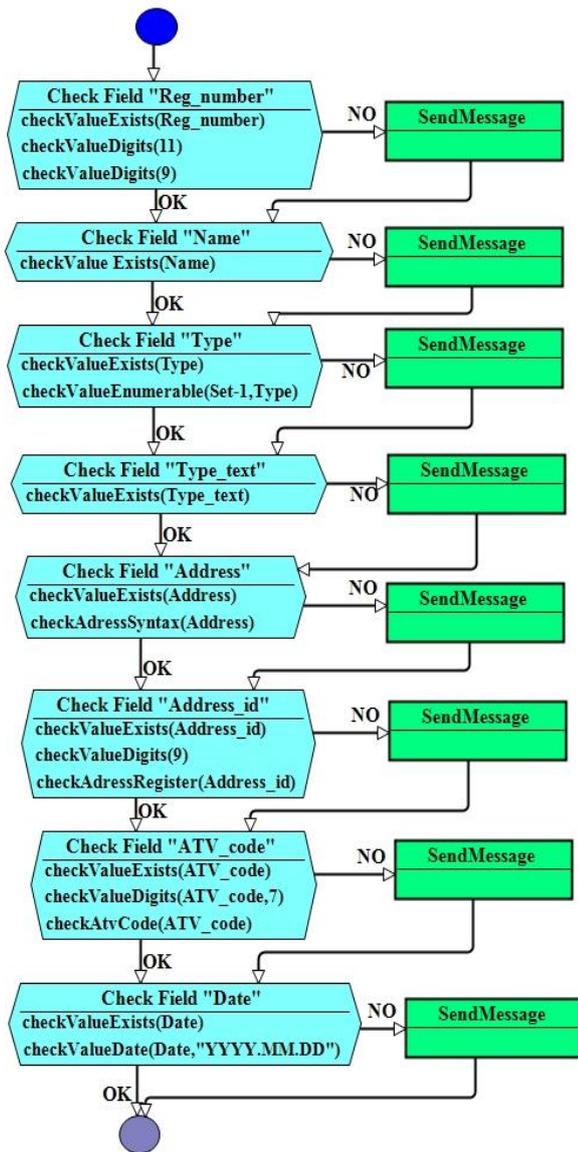


Fig.2. Quality specification of data object "Company"

In conclusion, steps to improve data quality can be performed by triggering changes in the data source.

The language describing the quality evaluation process involves verification activities for individual data objects which can be defined informally as a natural language text, or using UML activity diagrams, or in the own DSL. The Fig.3 contains separate field checks for the data object Company where each individual operation evaluates the data quality of the field by using a SQL statement. The SQL statement SELECT specifies the target data object, but WHERE specifies the quality specification. Such a data quality implementation is often used when data is stored in relational databases.

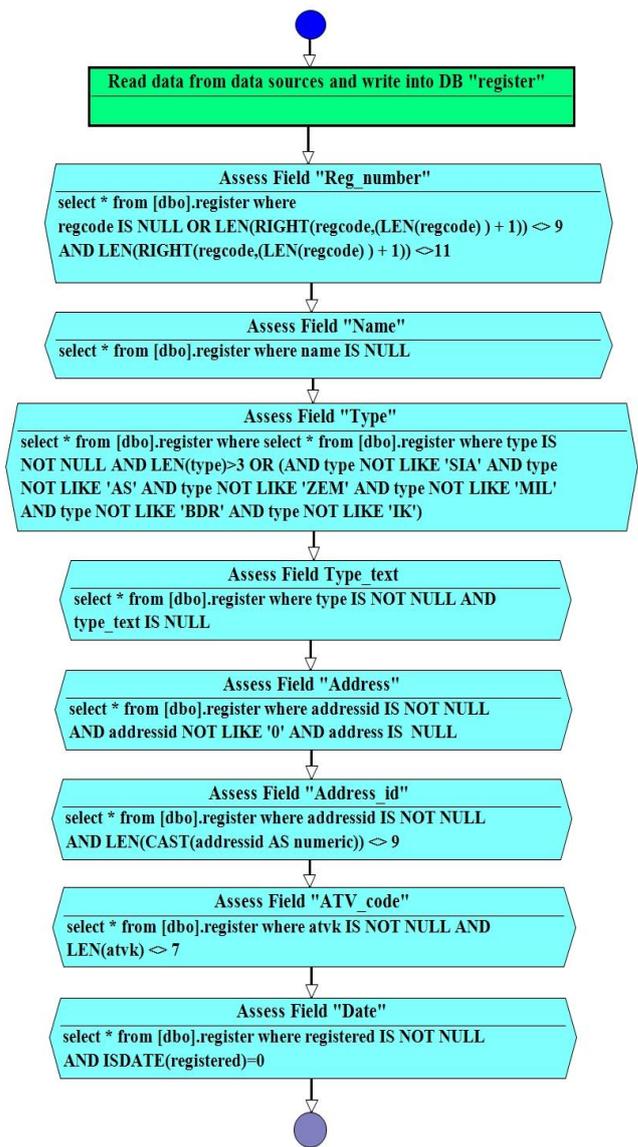


Fig.3. Quality evaluation process of data object "Company"

III. EXPERIENCE OF OPEN DATA QUALITY EVALUATION

The data quality evaluation methodology was approbated on real data sets analyzing open data published by company registers in four European countries (Latvia, Estonia, Norway and the United Kingdom). The quality of company registers' data was evaluated setting two quite simple tasks (use case): identification of all recorded companies and verification of recorded business addresses by contacting companies.

A company was identified using the following parameters: registration number, name, type of activity and registration date. In the Latvian Company Register, this information is stored in the data fields Reg_number, Name, Type and Date. If any of these fields was empty or did not correspond to the syntax rules, a company was not identified.

Company's business address with the postal code is needed to contact a company. The Company Register of Latvia stores

this information in fields Address and Post_code. If any of these fields is empty or does not correspond to the syntax rules, the company cannot be contacted via mail.

Open data of company registers of four European countries were analyzed to fulfill the task of company identification and contacting via mail for all companies registered in the particular country.

The conclusions were drawn that data quality evaluation depends on specific data use case.

a. Company Register of Latvia

The Company Register (CR) is a public register partly available as open data. The open data set of the Latvian CR [13] contains 396 thousand records. A company is selected as the data object for evaluation. The structure of this data object class partly (11 fields) is described in the Fig. 1 and Table 1.

Each company in the CR is characterized by 22 parameters. The data quality checks showed that 13 of 22 fields have no syntactic errors. But as it is shown in the Table 1, data quality problems were detected in 9 fields. NULL values of the field Name in 10 records and NULL values of the field Date in 94 records are considered as severe data quality problems. The company name and registration date is primary information about company and may not be left empty. If these records relate to real companies, then identification of them will not be possible.

TABLE I.
DATA QUALITY EVALUATION OF THE CR OF LATVIA

#	Field name	Field format	Error count	Comment
1.	Reg_number	11 digits NOT NULL	0	All companies have register number
2.	Name	NOT NULL	10 0.0025%	Company name is empty
3.	Type	NOT NULL Enumerable	0	Company type
4.	Type_text	NOT NULL Enumerable	1 403 0.35%	Company type (description) is empty
5.	Address	NOT NULL	366 0.09%	Company's legal address is empty
6.	Adress_id	9 digits, NOT NULL	4 523 1.14%	Company's address code is empty
7.	Region_code	9 digits, NOT NULL	280 662 70.7%	Region code is empty
8.	City-code	9 digits, NOT NULL	99 050 24.95%	City code is empty
9.	Post-code	4 digits, NOT NULL	20 498 5.16%	The postal code of company's address is empty
10.	ATV-code	7 digits, NOT NULL	5 521 1.39%	4 574 records - administrative territory code is empty 947 records - values are shorter than 7 digits
11.	Date	Date ('YYYY-MM-DD')	94 0.024%	Company's registration date is empty

However, if these records do not describe real companies, then these records should be removed from the CR. The number of incomplete records is not large and their processing would not require much work but this has not been done yet.

NULL values in the field Address in 366 records and NULL values in the field Post-code in 20498 records indicate potential data quality problems implying that this companies cannot be reached by mail. Other inaccuracies are not significant for the specific use case but they may be troublesome in other cases.

There are also 646 companies which according to their status are active but have NOT NULL value in "terminated" field which is not empty if only company's status is "closed" – liquidated or reorganized.

The blank values in other data fields, especially in the field Region_code, lead to the conclusion that CR and users obviously lack the specification of open data or have incomplete information about filling fields, have to interpret the meaning of fields and acceptable formats themselves. The different interpretations of data formats and content lead to data quality problems.

b. Company Register of Estonia

The data quality of Estonian CR [14] was analyzed using the data set of 266171 records with 14 fields for each record. The data of Estonian CR seems to be of higher quality than the data of Latvian ER. All fields identifying companies were filled in.

The registration date of a company is not included in the set of open data, so it cannot be used to identify a company. The identified data quality problems were NULL values in address field of 29918 records as well as NULL valued in other address-related fields (see Table 2). Values of fields Etevoja_address and ads_ads_oid are NULL in all records which suggest that the publication of values in these fields is unnecessary.

In addition to Ariregistri_kood there also exists a KMKR number - Estonian value-added tax identification. 178 550 records do not have any KMKR number, however the most part of these companies should have this number although the given data set does not indicate it. This means that the Register of Companies of the Republic of Estonia does not provide complete data about companies.

To summarize the data quality problems were detected in 7 of 14 fields.

Unlike the Latvian CR published data, the Estonian CR open data can be used to identify companies. Though contacting a company to business address by mail may be difficult due to blank address fields.

TABLE III.
DATA QUALITY EVALUATION OF THE CR OF ESTONIA

#	Field name	Field format	Error count	Comment
1.	Ariregistri_kood	8 digits NOT NULL	0	Company's registration number
2.	Nimi	NOT NULL	0	Company's name
3.	Ettevotja_staatus	Enumerable 'R', 'L', 'N'	0	Company's status is empty
4.	Asukoht_ettevotja_aadressis	NOT NULL	29 918 11.24%	Address is empty
5.	asukoha_ahak_tekstina	NOT NULL	19 964 7.5%	Text of address is empty
6.	indeks_ettevotja_aadressis	NOT NULL	22 621 8.5%	Company's address index
7.	ads_adr_id	NOT NULL	40 224 15.11%	Field is empty
8.	ads_normaliseeritud_taisaadress	9 digits, NOT NULL	40 099 15.1%	Field is empty
9.	ads_ads_oid			Field is empty
10.	ads_normaliseeritud_taisaadress	NOT NULL	40 099 15.1%	Field is empty
11.	Ettevotja_aadress			Field is empty

c. Company Register of Norway

The Norwegian CR [15] was analyzed using the open data set of 1 100 993 records with 42 fields. Data analysis shows that all fields identifying companies are filled in according to the formatting rules (see Table 3). Also, legal addresses are available for all companies. There are no postal codes in 14683 records which could be seen as a data quality defect. Also in other fields that should contain addresses of a company related activities have blank values. Similar as in registers of other countries, the value of the field *forretningsadresse_adresse* containing the legal address of the company is NULL in 68 128 cases.

Values of field *stiftelsesdato* containing date of company's liquidation are doubtful in 9 cases (1701-07-30, 1277-09-13, 1732-12-29, 1635-12-31, 1671-12-31, 1538-12-31, 1690-12-31, 1550-12-31).

TABLE IIIII.
DATA QUALITY EVALUATION OF THE CR OF NORWAY

#	Field name	Field format	Error count	Comment
1.	organisasjonsnummer	9 digits NOT NULL	0	Registration number of Company
2.	navn	Varchar NOT NULL	0	Name of Company
3.	Registreringsdato	date NOT NULL	0	Date of registration of Company
4.	hjemmeside	NOT NULL	0	Address of Company
5.	stiftelsesdato	date	9	End-Date of Company is doubtful
6.	forretningsadresse_adresse	NOT NULL	68 128 6.2%	Field is empty
7.	forretningsadresse_postnummer	Enumerable 'R', 'L', 'N'	22 362 (2%)	Field is empty
8.	forretningsadresse_poststed	NOT NULL	14 683 (1.3%)	Post address is empty
9.	forretningsadresse_kommunenummer	NOT NULL	22 362 (2%)	Field is empty
10.	forretningsadresse_kommune	NOT NULL	22 362 (2%)	Region address is empty
11.	forretningsadresse_landkode	NOT NULL	14 863 (1.3%)	Company's address code is empty
12.	forretningsadresse_land	9 digits, NOT NULL	14 863 (1.3%)	Company's country code is empty

In general, it must be acknowledged that the information used to identify company via company's business is properly addressed. However, information necessary for postal communication with the company is missing in some cases. In total, data quality problems were detected in 8 of 42 fields.

d. Companies House of United Kingdom

The CH of the United Kingdom [16] was analyzed using the open data set of 754 292 records which is about 20% of all registered companies.

Companies can be identified by values in fields *Company_number*, *Company_Name* and *IncorporationDate* (the date of foundation).

The quality of the data is very high, namely only one record does not have Company name and 3 records have dubious date values (see Table 4).

The stored addresses have quality defects since there are no addresses and/ or postal codes recorded for many companies.

TABLE IV.
DATA QUALITY OF THE CH OF UNITED KINGDOM

#	Field name	Field format	Error count	Comment
1.	CompanyNumber	Varchar (8) NOT NULL	0	All companies have register number
2.	CompanyName	Varchar (160) NOT NULL	1 NULL	Company Name is empty
3.	RegAddress.AddressLine1	Varchar (300) NOT NULL	7 514 NULL (1%) 4 invalid	Company Address (Lin-1) 4 records contain: "XXXXXXX", "XXXXXX", "XXXXXXXXXXXX", "XXX XXX" values
4.	RegAddress.PostCode	Varchar (20) NOT NULL	12 151 (1.6%)	Company Address Post code is empty
5.	CompanyCategory	Varchar (100) enumerable NOT NULL	0	
6.	CompanyStatus	Varchar (70) enumerable NOT NULL	0	
7.	URI	Varchar (47) NOT NULL	0	Format: http://business.data.gov.uk/id/company/X', where X - CompanyName
8.	Incorporation Date	Date (DD/MM/YYYY), NOT NULL	3	invalid values - "16/06/1701", "09/08/1638", "25/04/1552"

Less significant data quality defects were found for the indicated company addresses: different names are used for one country in the register in RegAddress_Country field (UNITED KINGDOM – 173 756 records and UK - 3; United States 447 and USA - 1, GREAT BRITAIN - 1, England - 5), certain listed values denote a particular area (WALES – 5 075, SCOTLAND, England & Wales – 184 097, England - 5, Virgin Islands - 41 and Virgin Islands, British - 22 and British Virgin Islands - 1) despite the fact that register contains countries value which unifies certain territories. Part of these values does not correspond with Companies House [14] policy which divides UK territory into Southern Ireland, England & Wales which companies are treated as a single entity and Scotland. Same tendency is observed in a country of origin: 464 records with country of origin of "Untied States" and 1 record – "United States of America"; 57 – "Great Britain", 3 – "England", 2 – "England & Wales" and 752 288 – "United Kingdom"; 143 – "Ireland", 9 – "Northern Ireland" and 2 – "Republic of Ireland"; 13 – "Nigeria" and 1 – "Republic of Nigeria". Moreover, 4 values does not corresponds with this field as they are not countries at all: "SW7" - South Kensington and part of Knightsbridge postcode, "EAST SUSSEX" which is a county in South East England, "BWI" - Baltimore/Washington International Thurgood Marshall Airport code, "DE 19901" Dover (city in the U.S. state of Delaware) postal code despite the fact that this field is supposed to contain only country names and there are specific

fields which are supposed to store postal codes and county names. There are companies from non-existing countries (Czechoslovakia, Jugoslavia, USSR which were registered after these countries have ceased to exist as political entities).

Moreover, there are 4 records where RegAddress AddressLine1, RegAddress AddressLine2, RegAddress PostTwown, RegAddress County and RegAddress Country values are "XXX".

In total, data quality problems were detected in 15 of 55 fields.

e. Results

This chapter analyzes the quality of data from company registers of 4 countries, which make some of their registry data open and available to public. A very simple example was chosen for data usage: (a) to search a company by its registration number or by its name, (b) once the company is identified, its address data is used to communicate with the company.

The performed data quality analysis is just one of the potential data usages. The data analysis was limited to the syntax analysis of data records, temporarily avoiding the analysis of interrelated (external) objects. As stated in [5], a deeper data quality analysis would lead to the analysis of links between records of the CR database and other data objects from external sources.

Despite the simplicity of the chosen data usage, inaccuracies were found in all four company registers. In total, percentage of columns with quality problems varies from 19% (in case of CR of Norway) to 50% (in case of CR of Estonia). Some quality problems could be easily solved thus improving the mentioned results significantly. Perhaps the authorities maintaining the registers are not even aware of this. This does not mean that the data from company registers cannot be trusted when a company must be identified since the number of defects is insignificant. The data in the company registers of Estonia and Norway can be used to identify companies without inconsistencies (all necessary fields are filled in). However, company register of Latvia (104 values - 0,0225%) and United Kingdom (4 values – 0,0005%) have several data quality problems which should be solved. However, correspondence with companies may fail, as the quality of address information is questionable. All analyzed registers had at least several data quality problems in the data fields containing the information be used contacting the company (address and postal code). The highest number of data quality problems were detected in the company register of Estonia (11,24% of address and 8,5% of postal code values were missing), the best results were shown by the company register of the United Kingdom (1% of address and 1,6% of postal code values were missing but also only 0,0005% of address values were invalid). These results show that data suppliers should inspect their data (authors of this paper recommend using the proposed approach) thus improving its quality. In addition, it can be concluded that small resources are needed to correct the few mistakes in the data to identify companies. But it would be much more difficult to complete address information.

IV. CONCLUSIONS

The paper has two main objectives: first, to verify authors' methodology of data quality evaluation using data quality models, and second, to determine whether or not the company register's published data are trustful, i.e., are these data of adequate quality for simple use. The study showed that:

- The data quality model allows describing the data quality independently of the information system that accumulates the data. Such an "external" data quality control mechanism enables the verification of "alien" data applicability within the context of our problem without the involvement of data holders.
- Using "alien" data from various sources of information, the author's had to reckon with the unprecise data definition in a meaningful sense. The data descriptions that were of interest to study were either not available or inaccurate. Thus, as in the case of company register, the authors had to interpret the data. This interpretation can be verified by analysing the data quality with the proposed methodology.
- The quality of the data can be described at least in two levels of abstraction: the one informally using natural language and another including executable program fragments or SQL sentences.
- Prior to use of the published by company registers in several countries, it is necessary to assess data reliability for a particular purpose. The open data quality analysis of the Latvian Company Register provided in this study serves as an example of data evaluation. The obtained results showed that the registers of all countries contain a small number of erroneous data. The most significant weakness of company data are fields describing company's legal or postal address; the values in these fields are just partially filled. In general, insufficient attention is paid to the quality of business company data. Even minor adjustments would make it possible to achieve significantly higher data quality.

This paper is a continuation of studies on information systems modelling [17], [18], [19]. Further research should focus on the creation of data quality models in the context of multiple data objects. It is also important to create a DSL to describe specific data quality activities.

ACKNOWLEDGMENT

This work has been supported by University of Latvia Faculty of Computing project AAP2016/B032 "Innovative information technologies".

REFERENCES

- [1] K. R. Popper, *The Open Society and its Enemies*, ([1945] 1966), 5th ed.
- [2] G.Arnica, J.Bicevskis, G.Karnitis, E.Karnitis: The mega-system: integration of national information systems : conceptual and methodological baselines . // Latvian Academic Library grey literature database. [Rīga], 2001
- [3] Redman, T.C.: *Data Quality. The Field Guide*, Digital Press, p. 74 (2001)
- [4] ISO 9001:2015: *Quality management principles*.
- [5] Bicevska, Z., Bicevskis, J., Oditis, I. Models of Data Quality. 12th Conference, ISM 2017, Held as Part of FedCSIS, Prague, Czech Republic, September 3-6, 2017, Extended Selected Papers. Lecture Notes in Business Information Processing, volume 311, pages 194-211 (2018)
- [6] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)* 41 (3), 16, 2009.
- [7] Freddie Bray, D. Max Parkin. Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness *European journal of cancer*. March 2009 Volume 45, Issue 5, Pages 747–755.
- [8] Carlo Batini, Monica Scannapieco. *Methodologies for Information Quality Assessment and Improvement. Data and Information Quality Dimensions, Principles and Techniques*. Springer International Publishing Switzerland 2016.
- [9] Data Quality Evaluation Methods. International SEMATECH Manufacturing Initiative. Technology Transfer #08074943A-ENG, 2008.
- [10] Steven Van den Berghe, Kyle Van Gaeveren. Data quality assessment and improvement: a Vrije Universiteit Brussel case study. 13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June. 2016, Scotland, UK.
- [11] Batini Carlo, Barone Daniele, Cabitza Federico, and Grega Simone. A data quality methodology for heterogeneous data. *International Journal of Database Management Systems (IJDMs-)*, Vol.3, No.1, February 2011.
- [12] Lewoniewski W. (2017) Enrichment of Information in Multilingual Wikipedia Based on Quality Analysis. In: Abramowicz W. (eds) *Business Information Systems Workshops. BIS 2017. Lecture Notes in Business Information Processing*, vol 303. Springer, Cham.
- [13] Company Register of Latvia. <http://dati.ur.gov.lv/register>
- [14] Company Register of Estonia. <https://opendata.riik.ee/en/dataset/http-avaandmed-rik-ee-andmed-ariregister>
- [15] Company Register of Norway. <http://data.brreg.no/oppslag/enhetsregisteret/enheter.xhtml>
- [16] Company House <https://www.gov.uk/government/organisations/companies-house>
- [17] J.Ceriņa - Bērziņa, J.Bičevskis, Ģ.Karnītis: Information systems development based on visual Domain Specific Language BiLingva, In: 4th IFIP TC2 Central and East European Conference on Software Engineering Techniques (CEE-SET), Krakow, Poland (2009).
- [18] J.Bicevskis, Z.Bicevska: Business Process Models and Information System Usability, *Procedia Computer Science* 77, pp. 72 – 79 (2015)
- [19] Bicevska, Z., Bicevskis, J., Karnitis, G.: Models of event driven systems. *Communications in Computer and Information Science* Volume 615, pp 83-98, (2016)

Evolution of the BPM Lifecycle

Marek Szelaǳowski
Vistula University
ul. Stokłosy 3,
02-787 Warszawa, Poland
Email: marek.szelaǳowski@dbpm.pl

Abstract—The process lifecycle systematizes the method of implementing and managing business processes in the organization. Due to changes in the social culture and the availability of technologies, the process lifecycle are also undergoing constant changes. The aim of this article is to analyze the direction of these changes and to propose a new process lifecycle, which would account for the requirements of the knowledge economy.

The article presents an overview of relevant literature on managing the process lifecycle. In the second part, it discusses changes to the principles of holding business operations, which are increasingly more limiting with respect to the scope of using traditional process management. In the third part, the article proposes an updated the business process lifecycle, which would adjust the lifecycle to observed business changes and make use of emerging ICT solutions. The proposed process lifecycle guarantees the coherence of the implementation process in KE.

I. INTRODUCTION

THE process lifecycle is a schematic overview of the method of implementing and managing processes in the organization. It has the role of a “map” of the fundamental business-process-managing process in the organization. Its role is to present the main idea, or the cardinal principles, of process management, in a manner which enables their coherent and intuitive understanding by those participating in the implementation at present or in the future. For this reason, the process lifecycle cannot be over-complicated, albeit it should nonetheless be detailed enough and practicable enough as to make possible the shift toward more detailed models, which capture in detail the workflow of specific stages of the process lifecycle in the organization.

Due to changes in the social culture and the availability of technologies, or, more generally speaking, changes to holding business operations, process management in general, and the process lifecycle in particular, are also undergoing constant changes. The aim of this article is to analyze the direction of these changes and to propose a new process lifecycle, which would account for the requirements of the knowledge economy and the development of ICT solutions, such as process mining, robotic process automation (RPA), machine learning (ML), and artificial intelligence (AI).

II. METHODOLOGY

The article presents an overview of relevant literature on managing the process lifecycle and on this basis puts forward a proposal of a more general approach to the process lifecycle in the organization; one taken from a process-centric perspective. In the second part, it discusses changes to the principles of holding business operations, which are increasingly more limiting with respect to the scope of using traditional process management and emerging ICT solutions. In the third part, the article proposes an updated model of the process lifecycle, which would adjust the lifecycle to observed business changes and make use of emerging ICT solutions, which offer real-time support to dynamic business process management.

III. THE BPM LIFECYCLE IN TRADITIONAL PROCESS MANAGEMENT

Literature presents numerous models of process lifecycles in the organization, which emerged within the framework of the traditional concept of business process management and were authored by:

- consulting and implementation companies, e.g. Gartner [1]
- software vendors, e.g. Software AG [2]
- academic researchers [3].

A. The process lifecycle

The concepts present illustrative approaches to the process lifecycle in the organization as a sequence of cyclical stages [4]. As a point of departure for this analysis the article selected the DMEMO cycle (an acronym coined from the first letters of the names of the subsequent stages: Design, Model, Execute, Monitor, and Optimize) [5], which is analogous to the DMAIC (Define, Measure, Analyze, Improve, and Control) cycle known from SixSigma [6].

Other process lifecycle models prepared within the framework of traditional business process management are also divided into stages presenting subsequent steps of the process lifecycle in the organization. Examples include:

- Define, model, simulate, implement, execute, monitor, analyze, optimize (Gartner) [1]
- Strategize, design, implement, compose, execute, monitor & control [2]

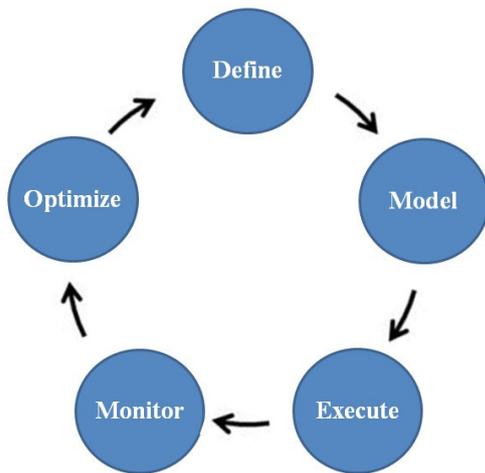


Fig. 1 The DMEMO process lifecycle
Source: [5]

- Model, simulate, implement, deploy & execute, monitor, optimize [7]
- Model, implement, execute, monitor, optimize [8]
- Discovery & remodeling, validation & simulation, deployment & execution, monitoring & performance management, improvement [9]
- (Re)design, configuration, enactment, diagnosis [10]
- Identification, modeling, implementation, controlling, process improvement [11]
- Analysis, design and modeling, implementation, monitoring and controlling, redefining and planning [12]

To generalize, within the framework of traditional process management the process lifecycle may be described as a cycle comprising sequentially executed stages with the aim of:

• Designing processes

This stage has the goal of preparing descriptions of processes existing in the organization (*as is*) and analyzing them on the basis of the organization's data, and, first and foremost, the knowledge of its personnel. In result of such analysis, an improved (*to be*) process model is prepared.

In traditional models that are commonly found in literature, this stage often contains or is defined as: identification, discovery, defining, redefining, designing, modeling, formalizing, simulation research, process optimization, etc.

• Implementing processes

This stage has the goal of accommodating the organization's operations to the designed process model. This accommodation encompasses both training and changes to the work of the personnel, as well as changes to the operations of the ICT infrastructure and the IT systems, including process performance automation.

In traditional models that are commonly found in literature, this stage often contains or is defined as: implementation, composition, positioning, process automation, etc.

• Process performance and monitoring

This stage has the goal of performing and monitoring business operations in accordance with prepared and implemented process descriptions. It is becoming increasingly more common in this stage to use techniques and analytical tools from the fields of BigData, internet of things (IoT), process mining, robotic process automation (RPA), machine learning (ML), artificial intelligence (AI), and expert systems.

In traditional models that are commonly found in literature, this stage often contains or is defined as: performance, monitoring control, measurement, etc.

• Process analysis and improvement

This stage has the goal of evaluating process performance and improving process descriptions with the aim of raising efficiency, minimizing risks, etc. At this point, techniques and analytical tools are used from the fields of BigData, process mining, artificial intelligence, and expert systems.

In traditional models that are commonly found in literature, this stage often contains or is defined as: analysis, diagnosis, optimization, improvement, etc.

B. *The life cycle of processes in the organization (BPM Lifecycle)*

Due to the identified necessity of approaching the process lifecycle from the perspective of implementing and performing multiple processes in the organization, the article proposes a process lifecycle model, which apart from the lifecycle of a single process also encompasses actions which from the perspective of the organization prepare the implementation of process management. This "global" life cycle of processes in the organization we will call Business Process Management Lifecycle in organization (in short: BPM Lifecycle). To this end, some consulting companies and researchers supplement the process lifecycle with an initial stage named:

- The formulation of vision [5]
- Process identification [3]
- Initial Process Planning and Strategy [4]

the aim of which is to define the goals and methods of process management in accordance with the strategy of the organization and its level of process maturity, prepare a corresponding plan of an implementation project for process management, as well as hold training courses for the organization's management and personnel.

This stage results in the preparation of a process architecture, which includes, among others, the agreed-upon goals and performance indicators, as well as priorities in the sequence of implementing particular groups of processes.

In traditional models that are commonly found in literature, this stage often contains or is defined as: planning, preparation, strategizing, identification, etc.

This elaboration, however, does not change the essence of depicting the process lifecycle (or BPM Lifecycle) within the framework of traditional process management as a sequence of stages performed one after another, preceded by a one-off execution of preparatory stages, which initiate the

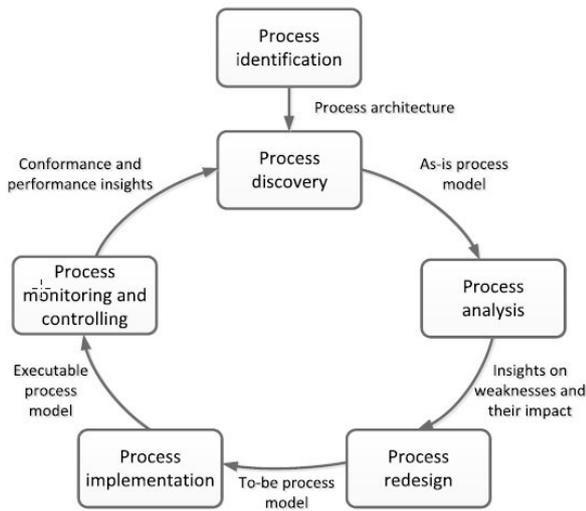


Fig. 2 The BPM lifecycle. Source: [3]

implementation of process management in the organization. For all practical purposes, it is an extension of E. Deming's PDSA cycle, which has been designed over 50 years ago, usually supplemented with additional “modern” ICT elements, such as: simulation, exploration, implementation, automation, reporting, etc.

IV. THE LIMITATIONS OF TRADITIONAL PROCESS MANAGEMENT

If we take a diagram depicting the BPM Lifecycle and replace symbols corresponding to subsequent stages (usually circles or ellipses) with symbols for subprocesses known from the Business Process Model and Notation (BPMN), the BPM Lifecycle (e.g. the model created by Dumas, La Rosa, Mendling, and Reijers from Fig. 2) will depict a normal, sequential “relay” process with a single feedback loop, the goal of which is to ensure periodical analysis and improve the process model on the basis of data derived in the course of its performance.

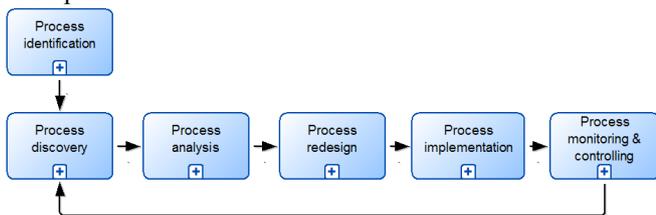


Fig. 3 The BPM Lifecycle as a process diagram in the BPMN notation. Source: Author's own elaboration, on the basis of [3]

In effect, processes cannot be improved upon or even changed at all during performance itself – this is possible upon completion and analysis alone. Most process-supporting workflow systems, document management systems, and Business Process Management Systems (BPMS) worked exactly in accordance with this principle. Even upon introducing changes to the course of a process, such changes will only be visible for process performances

which will be initiated after their acceptance (for new process instances). For processes which have already initiated performance (existing process instances) such changes are not visible. They are performed in accordance with an outdated version of the process description, even when it is apparent that it contains errors and may result in losses, and when we already know how the process may be improved upon. This nonsensical principle is further implemented in process-centric applications supporting process performance: the process performers use an application which was up to date in the moment of process initiation, even when an updated application is readily available.

This is fully in accordance with the principles of the traditional concept of process management, in which process performers are prevented from introducing changes in the course of performance itself. The course of the process is defined in the form of a description, or rather, an “algorithm,” prepared prior to initiating performance itself. In consequence, traditional process management lacks the possibility of quickly using knowledge obtained by the performers in the course of performance. In effect, this concept also does not offer the possibility of the operational use of new technologies, such as process mining, machine learning, or artificial intelligence, in the course of performance. Such use would require the authorization to change the process in the course of performance as the result of analyzing information obtained in the course thereof. This limitation results in the traditional BPM Lifecycle being inadequate in the case of about 70% of the processes performed in the knowledge economy [13][14]. This particularly pertains to essential processes, in which knowledge is constantly being created and verified, such as e.g. diagnostic-therapeutic processes, research and development processes, and personalized services.

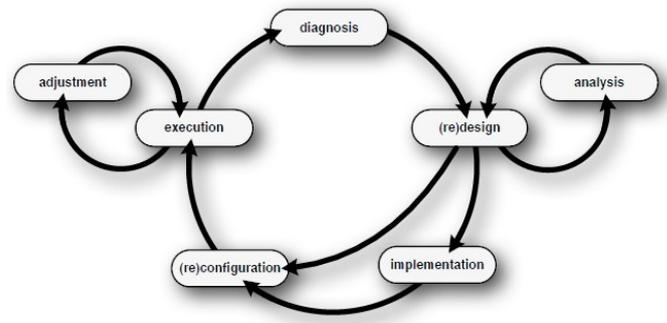


Fig. 4 The process lifecycle in the process mining methodology. Source: [15]

The management of such processes requires the process performers (or artificial intelligence) to be empowered to shape the processes on their own behalf, which requires introducing changes to the process lifecycle (and BPM Lifecycle), which are qualitatively deeper than just adding or subtracting subsequent stages of a sequential, routine cycle.

Among the process lifecycle models within the framework of traditional process management, a truly significant qualitative change was proposed in 2012 by the authors of the Process Mining Manifesto belonging to the IEEE Task Force on Process Mining (Fig. 4) [15].

The process execution stage has been supplemented with an additional "Adjustment" loop, the aim of which is to adapt the process in the course of performance itself. The standard process lifecycle has also been supplemented with a (re)configuration stage, in which changes are made to process-based executive systems (e.g. workflow management, document management, RPA, or BPMS) without having to repeat the implementation stage performed e.g. as the result of creating separate process performance scenarios. When recommending the change of the process lifecycle, the authors of the Process Mining Manifesto have stressed that organizations should also include the possibility of adjusting processes in the stage of designing processes and their supporting IT tools ("Analysis" loop). It has been clearly underlined that in the (re)design stage, analysis is held in the form of e.g. simulation research on the proposed process model or in the form of comparative analyses of the new process pattern with data on completed performances (researching compliance or extending the model as the result of process mining search) [13], with the end result being redesigned and reconfigured systems supporting process performance, e.g. RPA / using elements of AI or its integrating workflow systems / document management / BPMS.

This is a clear step toward changes to the process lifecycle, which allows for the dynamic management of processes. Having the option to improve processes in the course of their performance in the form of fixes, updates, adaptations, or limited experiments provides the process performers with the power to verify and create new knowledge in the course of their work with the use of machine learning or artificial intelligence. At the same time, the analysis of process performance in the (re)design stage allows for the uncovering of such knowledge thanks to process mining or analyzing the course of machine learning.

V. THE BPM LIFECYCLE IN DYNAMIC BPM

For full compliance with the concept of dynamic business process management, it is essential to manage the uncovered knowledge through the systemic combination of revealing knowledge with its evaluation and distribution. This, however, requires us to take the concept of process lifecycles in a direction in which the performance of a process will not be equal with the perfect repetition of the standard, but rather, the repetition or adaption of the standard with the best possible results in mind, in a manner which is the most adequate in a given context and within the limits of the executive privileges of the performer. Such adaptations may be introduced by:

- process performers

- process performers with the use of ICT solutions (e.g. online machine learning)
- elements of autonomic artificial intelligence

The postulated changes have been introduced in the process lifecycle model designed by the author in accordance with the concept of dynamic BPM. The model is presented on Figure 5.

The subsequent stages of the BPM Lifecycle of dynamically managed business processes are as follows:

Defining goals

In this stage, the goals of the project of implementing business process management, the goals of the megaprocesses, and the goals of knowledge management in the organization, as well as the principles of implementation themselves, are defined and agreed upon with the stakeholders.

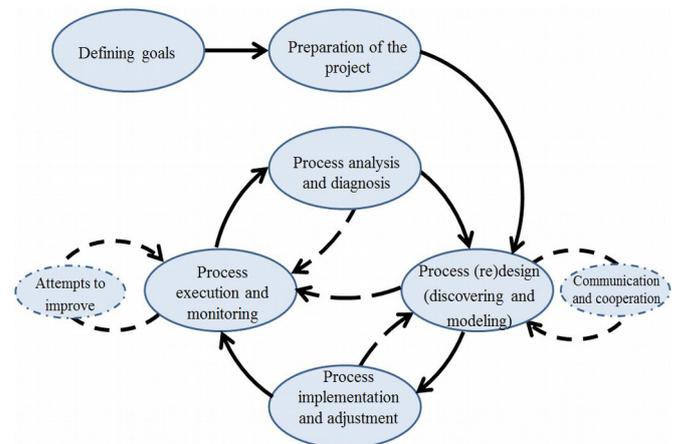


Fig. 5 The BPM Lifecycle in accordance with dynamic process management.

Source: author's own elaboration.

This stage results in the formulation of a definition of goals and a process map (and the de facto decision to initiate the implementation of business process management).

Preparation of the project

The goal of this stage is to prepare the organization for the implementation of process management by:

- defining or verifying the organization's level of process maturity[16]
- developing a method of process description and communication which is the most suited to the character of the performed processes
- holding training sessions for the management and the personnel of the organization.

The performance of this stage results in the creation of a process architecture and an implementation plan, which take into account the level of process and technological maturity and the culture of the organization.

(1) The (re)Design stage

In this stage, process descriptions and their corresponding data are created. Process discovery is performed with the use of:

- standard models for the field in which the organization operates,
- the knowledge of the employees,
- exploratory research (“discovery” / “mining”),
- analyses of data and the results of machine learning.

Depending on the level of dynamism of the processes involved, process descriptions may take the form of:

- for static processes – detailed descriptions, usually process models comprising interconnected process diagrams,
- for dynamic processes – detailed process descriptions in the form of models comprising process diagrams and/or collections of tasks to be accomplished during process performance (e.g. in the form of an ontology), as well as the data required during the decision-making process and in the documentation stage.

This stage should also encompass the preparation of requirements for RPA, as well as the preparation of prototypes of process-driven applications, which in the least should include the information content of the user interface, the possible range of standard reports, and the scope of integration with ICT infrastructure or BigData repositories. Furthermore, in this stage, the organization's internal rules and regulations should be – where required – updated for consistency between process management and other fields of management.

Communication and cooperation

In the (re)Design stage – in accordance with the principles of dynamic business process management – in order to make good use of the broadest possible part of the organization's intellectual capital, proposed process descriptions, prototypes, or applications and robotic process automation, which have been cleared for testing, should be consulted with in-house and external experts, and, first and foremost, with practitioners themselves, who use them on a daily basis, through e.g. communities of practice or social media websites.

(2) The Implementation and adjustment stage

In this stage, process descriptions are implemented (and eventual changes to other internal regulations are introduced) along with their supporting RPA and process-centric applications within the organization. In this stage, it is possible to adapt process descriptions and the configurations of their supporting robotic process automation and systems to the needs and requirements identified during implementation. Should it turn out that a designed process or configuration of a process-centric application does not meet the expectations of the users, it is possible to return to the (re)Design stage in order to prepare the process descriptions and applications once again.

(3) The Execution and monitoring stage

In this stage, business processes are performed and data on their performance is collected on an ongoing basis. For

transaction systems (e.g. MRPII, CRM, ERM, HIS, etc) and process-based systems (workflow / document management / BPMS), as well as RPA and AI, they are stored in event logs. Data from other sources (e.g. mobile applications, social media applications, e-mail accounts) should be integrated within a unified data source (BigData). Such information should be monitored by control systems on an ongoing basis, as well as analyzed and used in the ongoing support of knowledge workers by robotic process automation and/or elements of artificial intelligence.

Attempts to improve

In accordance with the 2nd principle of dynamic business process management [17], knowledge workers (and in the future – autonomic artificial intelligence as well) have the power to create or adapt described business processes to the requirements of a specific context of performance and the changing general conditions of process performance (e.g. changing technologies, principles of competition, or the individual, unpredictable context of performance). Such active experiments have the goal of arriving at new solutions enabling the performance, or the optimization of the performance, of a process.

(4) The Analysis and diagnosis stage

In addition to business processes being monitored in the Execution and monitoring stage, business processes are nevertheless evaluated ex-post by means of:

- standard control actions, including the control of process efficiency, duration, costs, resources used, risks involved, etc.;
- discovering the actual course of the performed processes and evaluating the results of the implemented improvements with the aim of:
 - broadening the processes of the organization through communication (adding to the list of best practices and informing about the update), as well as redesigning and tailoring processes and their supporting applications and robots;
 - communicating information on the negative results of specific attempt at improving a process (adding to the list of wrong practices and informing about the update);
 - initiating a broader evaluation of the possibilities of using a discovered potential improvement (while informing the stakeholders about the possibility of participating in the discussion).

Knowledge obtained in this stage should be systematically communicated to authorized members of the organization, with a particular focus on the employees who are directly responsible for process performance, for whom new or verified knowledge might have direct significance (in the Execution and monitoring stage). This requires the existence within the organization of a culture and mechanisms of internal communication, which allow for the ongoing, broad improvement of processes and the distribution of knowledge, as well as the existence of an ICT

infrastructure enabling the rapid introduction of changes and their communication.

At the same time, within the proposed lifecycle model for dynamically managed processes improvements resulting from practical attempts at innovation, which have been given a positive evaluation, may be introduced in the Execution and monitoring stage directly following the (re)Design stage, without the necessity of going through the Implementation and adjustment stage. As previously, this requires organizations to develop efficient mechanisms of internal communication both on the level of social culture, as well as on the level of ICT infrastructure, understood as e.g. the broad acceptance and the efficient use of mobile devices, social media applications, or elements of artificial intelligence.

VI. CONCLUSION

In the knowledge economy, a mere 30 percent of processes within the organization are static in nature, for which detailed models or even algorithms may be prepared prior to performance [18][19][20]. The remaining 70 percent of processes are processes which require dynamic management, or the empowerment of their performers to introduce changes in the course of performance itself. As has been shown in the article, the development of process management requires the introduction of a qualitative change to the process lifecycle, which would account for the possibility, and in the case of a large majority of dynamic processes – the necessity, of using the knowledge of the process performers to tailor the processes to the context of a specific performance. Without this change it is impossible to make efficient use of new technologies, such as process mining, machine learning, or artificial intelligence. Within the framework of traditional process management, the use of such technologies in the course of a process lifecycle is impossible or ineffective, as it provides benefits only upon subsequent performance of the process in question (or upon an even more delayed approval of the change by a group of process owners). In the knowledge economy, implementations of process management in accordance with the traditional BPM Lifecycle were seen and remain to be seen as a success only because:

- they pertain to static (repeatable, routine, unchangeable) processes, the optimization or automation of which (e.g. through RPA) allows us to raise the pace of performance and lower costs and risks
- sub-optimal performance or losses during performance are so high that in effect any improvement initiatives bring about tangible effects

However, the situation is changing due to:

- the number of static processes in the organization steadily becoming lower
- the possibility of using new ICT technologies, among which one should primarily mention those which

work in real time: process mining, machine learning, and artificial intelligence.

Taken together, both these factors result in the scope of processes requiring dynamic management becoming larger, and, at the same time, allow access to a growing number of tools supporting knowledge workers in this regard. Nevertheless, they exert growing pressure on the organization on the part of the competition and the clients. The BPM Lifecycle proposed in this article requires the adjustment of methodologies and tools supporting process management with a view to the efficient use of both emerging ICT technologies and the intellectual capital of the organization, encompassing the entire process lifecycle.

VII. REFERENCES

- [1] G. Polancic, "Learning BPMN 2.0 – Business Process Vs Workflow," 2013, <http://blog.goodelearning.com/bpmn/business-process-vs-workflow/>
- [2] Software AG, "Enterprise BPM series: a summary," 2011, <http://www.ariscommunity.com/users/nina-uhl/2011-07-26-enterprise-bpm-series-summary>
- [3] M. Dumas, M. La Rosa, J. Mendling, and H. Reijers, *Fundamentals of Business Process Management*. Heidelberg: Springer, 2016.
- [4] R. Macedo de Morais, S. Kazan, S. Dallavalle de Padua, and A. Costa, "An analysis of BPM lifecycles: from a literature review to a framework proposal," *Business Process Management Journal*, May pp. 412-432, 2014, doi: 10.1108/BPMJ-03-2013-0035.
- [5] BPM Resource Center, "Understanding BPM and Related Improvement Methodologies," 2014, http://www.what-is-bpm.com/get_started/bpm_methodology.html
- [6] P. Pande, R. Neuman, and R. Cavanagh, "Six Sigma", Warszawa, K.E. Liber S.C., 2003, pp. 37, 141-142.
- [7] T. Gullidge, S. Hiroshige, and D. Manning, "Composite Supply Chain Applications," 2011, <https://www.intechopen.com/books/supply-chain-management-new-perspectives/composite-supply-chain-applications>, p. 49, doi: 10.5772/21927.
- [8] PNM SOFT, "BPM Lifecycle," 2017, <http://www.pnmsoft.com/resources/bpm-tutorial/bpm-lifecycle/>
- [9] A. Pourshahid, D. Amyot, L. Peyton, S. Ghanavati, P. Chen, M. Weiss and A. Forster, "Business process management with the user requirements notation," *Electronic Commerce Research*, vol. 9, 2009, pp. 269-316, doi: 10.1007/s10660-009-9039-z
- [10] C. Di Ciccio, A. Marrella, and A. Russo, "Knowledge-intensive Processes Characteristics, Requirements and Analysis of Contemporary Approaches," *Journal on Data Semantics*, vol. 4, no. 1, pp. 29-57, https://www.researchgate.net/profile/Claudio_Di_Ciccio/publication/269629902_Knowledge-intensive_Processes_Characteristics_Requirements_and_Analysis_of_Contemporary_Approaches/links/576a501a08ae1a43d23bca3c.pdf
- [11] A. Bitkowska, „Zarządzanie procesowe we współczesnych organizacjach,” 2013, Warszawa: Difin SA, pp. 66-83.
- [12] R. Bernardo, G. Ribeiro, and S. Dallavalle de Pádua, "The BPM lifecycle How to incorporate a view external to the organization through dynamic capability," *Business Process Management Journal*, vol. 23, no. 1, 2017, pp. 155 – 175.
- [13] E. Olding and C. Rozwell, "Expand Your BPM Horizons by Exploring Unstructured Processes.", 2009, Gartner Technical Report G00172387,.
- [14] K. Swenson, "Mastering the Unpredictable: How Adaptive Case Management Will Revolutionize the Way That Knowledge Workers Get Things Done." Tampa (USA): Meghan-Kiffer Press. 2010. pp. 30-31.
- [15] Process Mining Manifesto. 2012, http://www.win.tue.nl/ieeetfpm/doku.php?id=shared:process_mining_manifesto
- [16] K. Kania, „Doskonalenie zarządzania procesami biznesowymi w organizacji z wykorzystaniem modeli dojrzałości i technologii informacyjno-komunikacyjnych,” Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, 2013, pp. 85-87.
- [17] M. Szelągowski, "Becoming a Learning Organization Through Dynamic Business Process Management," *Journal of Entrepreneurship*,

- Management and Innovation (JEMI)*, vol. 10, no. 1, 2014, pp. 147-166.
- [18] HandySoft, "Dynamic BPM – The Value of Embedding Process into Dynamic Work Activities: A Comparison Between BPM and Email." 2012, http://www.bizflow.com/system/files/downloads/HandySoft%20-%20Dynamic%20BPM%20White%20Paper_0.pdf.
- [19] T. Austin, "Gartner Says the World of Work Will Witness 10 Changes During the Next 10 Years." 2010, <https://www.gartner.com/newsroom/id/1416513>
- [20] J. Ukelson, "Adaptive Case Management over Business Process Management." 2010, <http://it.toolbox.com/blogs/lessons-process-management/adaptive-case-management-over-business-process-management-40002>.

Software Systems Development & Applications

SSD&A is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the discipline of software engineering. The SSD&A area emphasizes the issues relevant to developing and maintaining software systems that behave reliably, efficiently and effectively. This area investigates both established traditional approaches and modern emerging approaches to large software production and evolution. Events that constitute SSD&A are:

- MDASD'18—5th Workshop on Model Driven Approaches in System Development
- MIDI'18- 6th Conference on Multimedia, Interaction, Design and Innovation
- LASD'18—2nd International Conference on Lean and Agile Software Development
- SEW-38 & IWCPS-5—Joint 38th IEEE Software Engineering Workshop (SEW-38) and 5th International Workshop on Cyber-Physical Systems (IWCPS-5)

AREA SUPERVISORY COMMITTEE

- Hinchey, Mike, SEW-38-IWCPS-5
- Kornecki, Andrew J., MMAP'18
- Luković, Ivan, MDASD'18
- Marasek, Krzysztof, MIDI'18
- Przybyłek, Adam, LASD'18

6th Conference on Multimedia, Interaction, Design and Innovation

MIDI Conference provides an interdisciplinary forum for academics, designers and practitioners to discuss the challenges and opportunities for enriching human interaction with digital products and services.

The main focus of MIDI Conference is exploring design methods for creating novel human-system interaction, developing user interfaces and implementing innovations in user-centred development of advanced IT systems and on-line services.

TOPICS

Topics of interest include (but are not limited to) the following areas:

- interactive multimedia and multimodal interaction design
- novel interaction techniques, voice interfaces, interactive multimedia
- ubiquitous, multimodal, pervasive and mobile interaction, wearable computing
- novel information visualization and presentation techniques, Augmented/Virtual Reality
- design methods for usability, accessibility and outstanding user experience
- prototyping of user interfaces and interactive services
- human-centred design practices, methods and tools, user interface design
- unfolding trends in HCI research and practice, customer experience, Service Design
- advances in user-centred interaction design
- understanding people and interactions: theory, concepts, models and methods
- understanding people and interactions: contextual, ethnographical and field studies
- critique and evolution of methods, processes, theories and tools for human-computer interaction
- novel methodologies for conceptualization, design and evaluation of interactive products and services

EVENT CHAIRS

- **Marasek, Krzysztof**, Polish-Japanese Academy of Information Technology, Poland
- **Romanowski, Andrzej**, Lodz University of Technology, Poland

- **Sikorski, Marcin**, Polish-Japanese Academy of Information Technology, and Gdansk University of Technology, Poland

PROGRAM COMMITTEE

- **Biele, Cezary**, Information Processing Institute, Poland
- **Brocki, Łukasz**, Polish-Japanese Academy of Information Technology
- **Forbrig, Peter**, University of Rostock
- **Grudziński, Krzysztof**
- **Guttormsen, Sissel**, University of Bern, Institute of Medical Education, Switzerland
- **Kaptelinin, Victor**, Umea University
- **Korżinek, Danijel**, Polish-Japanese Academy of Information Technology, Poland
- **Kołakowska, Agata**, Faculty of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
- **Landowska, Agnieszka**, Gdansk University of Technology, Poland
- **Marti, Patrizia**, University of Siena, Italy
- **Masoodian, Masood**, Aalto University
- **Miler, Jakub**, Faculty of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
- **Pribeanu, Costin**, National Institute for Research and Development in Informatics - ICI Bucuresti
- **Satalecka, Ewa**, Polish-Japanese Academy of Information Technology
- **Slavik, Pavel**, Czech Technical University
- **Wichrowski, Marcin**, Polish-Japanese Academy of Information Technology, Poland
- **Wieczorkowska, Alicja**, Polish-Japanese Academy of Information Technology, Poland
- **Winkler, Marco**, University Paul Sabatier
- **Wojciechowski, Adam**, Institute of Inf. Techn., Lodz Univ. of Techn.
- **Woźniak, Paweł W.**, University of Stuttgart, Germany
- **Wołk, Krzysztof**, Polish-Japanese Academy of Information Technology, Poland
- **Ziegler, Juergen**, University of Duisburg-Essen

Use of fuzzy cognitive maps for enhanced interaction with multiple mobile devices

Przemysław Kucharski, Dawid Sielski, Tomasz Jaworski, Andrzej Romanowski, Jacek Kucharski
Institute of Applied Computer Science
Lodz University of Technology
ul.Stefanowskiego 18/22, 90-924 Łódź, Poland
Email: pkuchars@iis.p.lodz.pl

Abstract—The aim of this work was to design and implement a mechanism supporting collaborative sensemaking in a system of multiple mobile devices implementing spatial awareness. The design is based on an observation how people tend to manage and organize information in physical space. The developed mechanism attempts to determine the relation between atomic elements of information basing on relative position of these elements in time. The proposed solution is based on a simple Fuzzy Inference System and the theory of Fuzzy Cognitive Maps. Physically, the system was implemented for three tablets, for which spatial awareness is simulated with the use of motion tracking system. The system was evaluated in a user study.

I. INTRODUCTION

IN the modern world, human beings and technology have to coexist on daily basis. One of the aspects of this coexistence is how people perceive information using technology. As life is getting faster, the amount of information reaching a person is increasing. Due to that, the mechanisms used to absorb all this information have to become more and more effective. When people are facing information, they have to somehow perceive and understand what this information means - this process is called sensemaking, from "making sense" of information. There exist many techniques to support this process - one of most common examples is underlining the important parts in a text - it simplifies cognition and hierarchizes the information. Another habit many people have, a technique that underlies the following work, is using physical space to support sensemaking - for instance, a student learning for an exam puts the notes in a certain order on the floor, creating a physical space of knowledge - they may guess what each pile of notes contains basing on where it is on the floor. As it was stated, nowadays people use technology, e.g. mobile devices as smartphones and tablets, to explore data and deal with information. There is an emerging need to create technological solutions that will be able to support data exploration processes such as sensemaking. This work is another step forward to obtaining solutions that will help people understand information they see. It was initially inspired by the pace of experiments in this field, and is basing on the earlier step in the work [1] a multi-device, spatially-aware interface that was designed to support sensemaking. The goal of this project is to design and develop mechanism for analysis of spatial data management in multi-device system. The mechanism is based on a complete weighted graph. The design is inspired by the theory of Fuzzy

Cognitive Maps (FCM). The graph reflects mutual relations between atomic elements of information in a set of hints given to solve a problem.

A. Motivation

The pace of world development poses new challenges in the field of Human-Computer Interaction. People nowadays use mobile devices at all times, and it is not strange to spot a person carrying more than one of them, for example a tablet and a smartphone. It is also more frequent to use mobile devices during meetings, so there is a need for solutions in the field of collaborative sensemaking. The previous work, focusing on the use of spatial awareness in an interface for mobile devices, gave promising results for further development [1]. In the course of work in this area, an idea of implementing the methods of artificial intelligence arose.

II. RELATED WORK

The following work covers the areas of data exploration, in particular collaborative sensemaking. It also benefits from the preceding work and insights in the fields of multi-device environments and spatially-aware systems. The original approach, consisting of implementation of AI methods, bases on the theory of Fuzzy Cognitive Maps.

A. Data exploration

As stated before, technological development requires addressing new challenges in the field of gaining knowledge and understanding the information. One of important issues is how to answer the situation of data exploration in everyday situations. As one of the aspects of ad-hoc data exploration, Fjeld et al. [2] raised the issue of "big data" present in public space - which means that the load of information in this area is rapidly increasing. What lays in connection to that, a considerable part of this information load may be inaccessible to most people. They considered possible designs for ad-hoc data exploration and proposed tangible tabletops as one of possible solutions, with an insight of future technology development for interactive environments. However, this solution assumes introduction of a new equipment to public spaces, where they can serve as data exploration tools. On the other hand, Weise et al. [3] investigated how the issue of data administration, exploration and analysis should be addressed in the age of

ubiquitous computing. They suggested connecting it to the local infrastructure, both in terms of human awareness and environmental sensing.

B. Multi-device environments

As it was mentioned, situations when users have several mobile devices to work at the same time can often be encountered. Currently available applications, however, do not provide solutions by which users can benefit from the interactions between the devices they use, i.e. cross-device interactions. The subject of multi-device environments was considered for a long time. Bilezikjian et al. [4] explored how the interactions with handheld devices may look like, although they were not commercially available. Blackwell et al. [5] investigated the issue of tangibility in the context of mobile devices. Cauchard et al. [6] discussed the concerns and opportunities for visual aspect of mobile multi-display environments. In Conductor [7], Hamilton and Wigdor presented a framework for examining the scenarios of cross-device interactions. They provide functionalities to split the aspects of performed task between several mobile devices. They also elaborated possible usage scenarios for the system, demonstrating their way of understanding cross-device interactions in task-specific domain. The study performed using the developed multi-device system demonstrated that using several connected devices is highly useful to perform certain tasks. Most of the participants made use of multiple devices and of the functionalities enabling them to easily transfer information across devices. The Pass-them-around [8], [9] system developed by Lucero et al. showed that providing people with the functionality of sharing content between their personal devices may lead to enrichment of interactions between users. Cassens et al. [10] proposed a taxonomy for the term cross-device interactions. They introduced the dimensions of ownership, distance and access to classify those interactions. They discussed this classification in the context of the work published in this field. Finally, they proposed a definition: "Cross-device interaction (XDI) is the type of interaction, where human users interact with multiple separate input and output devices, where input devices will be used to manipulate content on output devices within a perceived interaction space with immediate and explicit feedback." In general, the past research in the field of multiple device ecologies demonstrates that once cross-device interactions are implemented and enabled, users adapt to the new environment and benefit from these interactions. This proves, as it was raised in many of the papers, that further exploration of this field is necessary and promises a valuable contribution to our future life.

C. Spatially-aware systems

Space and awareness of space is of great importance for human perception of surrounding world. Hall developed a theory describing the relation between spatial arrangement of people and their social behaviours and emotions. Chen and Kotz [11] included spatial awareness as an important aspect of

context-aware computing. Some past research show that space may also have a vital impact on understanding and learning processes [12]. There was also some investigation concerning how transferring the perception of space to technology may help users interact with data.

In MochaTop [13], WoÅźniak et al. made a step forwards to understand how spatial combinations of two devices can be used for data exploration. They focused on a single user scenario for ad-hoc interactions, basing on an assumption that a couple of a smartphone and a tablet is more and more often carried by users. They designed and implemented several spatial-based interactions for exploring complex structures of data.

Their study showed that users find it useful to make space one of the input sources in data manipulation. Spindler [14], [15], [16] proposed a way of extending the interaction space into the third dimension. In this work, users are in disposition of several spatially-aware tangible displays to interact with virtual objects present on a central tabletop. This work demonstrated the potential of employing explorable 3D space for visualizations. By direct translation of spatial position to visualization input, he created a tool for intuitive exploration of several types of complex data sets.

The field of extending the interaction space of one mobile device was also investigated. AD-binning [17], which is an abbreviation from Around-Device Binning, is a mobile user interface that allows users organize data spatially outside of the device space. The concept consists of creating virtual zones (bins) around the device. The system tracks relative position of user's finger and the device and thus makes it possible to put pieces of information from the screen into these zones. The motivation of designing such solutions is, as stated by the authors, extending the interaction space due to insufficiency of that space on the screen of a mobile device. The premise that the design of user interfaces for multiple mobile devices should be based on spatial awareness of those devices was discussed by Rädle et al. [18]. They divided cross-device interactions into three groups: (i) spatially agnostic, (ii) based on synchronous gestures and (iii) spatially-aware. Spatially agnostic interactions may be menu-based with some kinds of user-perceivable device identification. Interactions based on synchronous gestures are triggered by simultaneous interaction with two or more devices. Eventually, spatially-aware interactions are based on relative arrangement of devices. The study conducted to compare them demonstrated that in most cases, spatially-aware interactions between devices are expected by the users. However, similarly to the Ballendat's conclusion about proxemics, Rädle et al. clearly emphasise that the way in which spatially-aware interactions should be designed is not yet fully explored and requires further investigation.

1) *Hardware for spatial awareness:* There are many attempts to find technical solutions for spatial awareness of mobile devices. There are approaches that may benefit both the further examination of spatial interactions, as well as those which may be applied in real world.

HuddleLamp [19] provides a solution for creating ad-hoc

multi-device ecologies on a tabletop based on video processing. Rädle et al. propose spatially-aware multiple device systems as an alternative for interactive tabletops. An interesting approach in this work is tracking not only the devices, but also the motion of users' hands to enrich interactions. One of technical possibilities that may be implemented in commercially available devices was presented by Ellyptic Labs, with a ultrasound Doppler-based system for spatial awareness of mobile devices.¹ Another emerging direction is embedding the sensing in the environment. An effort was made recently to develop tomography-based table for spatial positioning of devices [20]. This solution is based on the change in electrical capacitance triggered by mobile devices, which usually consist of magnetic elements. This approach gives promising results from the very beginning, yet it is in early stage of development.

D. Fuzzy logic

Fuzzy logic is a form of mathematical logic where the variables may not only be 0 or 1, which is true or false, but may be any real number from this interval[21]. This makes it possible to mathematically introduce the concept of something being partially true. For linguistic variables, a membership function can be introduced to manage the degree of truth[22]. This membership functions characterize various sub-ranges of a continuous variable. By the process of fuzzification, it is possible to map the input value of a variable to its membership function, and therefore obtain a more accurate result than with the use of crisp logic. The concept of fuzzy logic is used in many applications. For instance, Zadeh [23] proposed a methodology for computations using very imprecise information, where words are used instead of numbers. This approach requires the use of fuzzy logic and fuzzy set and promises very effective solutions for many problems. Another very interesting work in the field of fuzzy logic and sets, fitting closely to the scope of this project, is the theory of fuzzy information granulation proposed by Zadeh [24]. This work elaborates the human reasoning system in the point of view of automatization using fuzzification and granulation.

E. Fuzzy cognitive maps

The concept of Fuzzy Cognitive Maps was proposed by Kosko [25] as a structure to represent causal reasoning. The theory is based on an observation that in knowledge processing, most of the relations, including classification and causality, are uncertain and imprecise. Fuzzy Cognitive Maps are graph structures, where the nodes of the graph represent causal objects or concepts, and the edges represent mutual relation between these concepts. The concept of FCM is based on the theory of cognitive maps introduced by Polish scientist Axelrod [26] for representing social scientific knowledge. Papageorgiou [27] stated that "Fuzzy cognitive maps (FCMs) are a modeling methodology based on exploiting knowledge and experience." This concept has been successfully applied in many fields, including political decisions [28], modeling

electrical circuits [29] or organisational behaviour [30]. Most of the implementations of FCMs require expert in the process of map design, but Aguilar reveals in his survey [31] that there are attempts to develop FCMs automatically from raw data. This survey also shows a wide range of possible applications, covering both causal and non-causal relationships with imprecise information.

III. THEORETICAL BACKGROUND

The design and implementation of the proposed mechanism poses several challenges. The first is the determination of mutual position of two atomic elements of information in the space. This is possible thanks to the motion tracking system, which sends the position of each device configured as a rigid prop. The information given by the system is translation versus the origin of global coordinate system and the rotation in Euler angles. On the other side, the device sends the position of a post-it on its screen. Basing on this, a matrix calculus enables determining the position of each post-it in the global coordinate system.

The most important part is the model of the data set, containing information about the mutual relation between set elements. A concept responding well to the needs of this project is Fuzzy Cognitive Map. Therefore, the entire data set was modeled in the system in a form of complete weighted graph. Some implementations of FCM are used to describe causal relations, and for this purpose directed graphs are used. In the case of this design, the aim was simply to reflect the strength of mutual relation between each two elements – the atomic elements of information are represented as graph nodes, and the weight of the edge connecting two nodes represents the aforementioned strength of relation.

In this model, it was also extremely important to reflect the imprecise nature of the relations. This was obtained by using FCM, not the original concept of Cognitive Maps. In the model, the weight of the node is any number from 0 to 1, where 0 – no relation, 1 – strong relation. This makes it possible to show the user not only which elements are related, but also give them hints which of them are related stronger than others.

IV. MODELLING THE RELATIONS

The class InteractionModel manages all the position data concerning both the devices and the resources in form of textual hints. The original approach in using this positional information is used to determine the numerical values in the weighted graph. A complete weighted graph containing 31 nodes is implemented in form of an array. The positions of each resource in global coordinate system are updated basing on data from tablets and motion tracking. In order to do that, Java library used for matrix calculus is used. The server calculates the distance between the resources. The distance is normalized to scale 0-10 basing on maximum distance in the current step. Then, the normalized distance is passed to mechanism that uses fuzzy logic. To implement fuzzy logic rules, a Java library jFuzzyLogic [32], [33] was implemented

¹www.ellypticlabs.com

in the project. It uses Fuzzy Control Language to implement the fuzzy system and a interface to use the system in Java. The output variable, which is the change in relation, is returned after evaluation of the system. The weights in the graph are changed according to the response of fuzzy system, and then the weights in the entire graph are normalized to interval 0-1, basing on the maximum and minimum value of relation found in the graph. If the user chooses to display relations of one of the post-its by clicking it, the InteractionModel returns the related post-its (i.e. which have relation strength with the chosen one different from 0) along with the weight of the graph edge.

V. STUDY AND RESULTS

The method of evaluation of the developed solution is a user study. User study is a well known method in Human-Computer Interaction. It is a way to experimentally test a hypothesis. It can test a variety of measures, for example user experience, efficiency, accuracy, in general both qualitative and quantitative performance of a system. There are several procedures of planning and running user studies [34]. In this project, two systems are tested to determine the potential difference in their performance from the point of view of a user:

- A system supporting static relations between elements
- A system supporting adaptive relations between elements based on interaction of the user

A. Experimental procedure

The hypothesis of the study is that the system offering adaptive relation feedback may have a positive impact on the effectiveness of the solution. The study is divided into two parts: in one, the study apparatus is the simplified version of the system with static relations to have baseline results, in the other one, current implementation is used in order to measure the differences. The study in each case consist of several parts. In both stages, the system was set up in an isolated environment. Participants were informed on the purpose of the study and instructed on the task and how to use the system. Then they were asked to solve the mystery. The time for solving the mystery was approximately 30 min. After that, they were shortly debriefed on the functionalities of the system and overall opinion about the proposed solution.

B. Task description

The task in the study was to solve a crime mystery. In the data set, there were 31 hints containing information about places, events and people. The questions that were to be answered included: the person of murderer, the time of murder, the place and the motive. The preliminary fact is that Mr Kelley was murdered. The questions to be answered involve the person who killed, the place, time and the motive of murder. The structure of the task is complicated: there are several people, some of which have nothing to do with the crime under investigation. The evidence is uncertain and the information about the place inconsistent. To show the difficulty

of the task, it is enough to quote the prepared answer: "After receiving a superficial gunshot wound from Mr. Jones, Mr. Kelley went to Mr. Scott's apartment where Mr. Scott killed him with a knife at 12:30 AM because Mr. Scott was in love with Mr. Kelley's wife". The study task was based on a criminal mystery from a book prepared for teaching collaboration skills [35].

C. Study results: static relations system

In the first stage of study, with the version of the system supporting static relations established among atomic information elements, $n = 16$ participants in 8 pairs (aged 24-61, $\mu=37$, 9 males, 7 females) were asked to complete the task (Figure 1). They completed the task in mean Task Completion Time (TCT) of 21 minutes and 18 seconds, with standard deviation of 3 min 55 sec. At this stage, the relations in the system were predetermined, basing only on text analysis of the hints. The system supported highlighting related hints and displaying a time line - an ordered connection between hints containing temporal information.



Fig. 1. Static relations version of the system

D. Study results: adaptive relations system

The second stage of the study was performed with the use of current implementation of the system. In this study, $n = 8$ participants in 4 pairs (aged 15 – 23, $\mu = 21$, 5 males, 3 females) were asked to solve the crime mystery (Figure 2). For the purpose of this stage, the same crime mystery with exactly the same hints was used. Participants came with the proper response in mean time of 19 minutes 41 seconds, with standard deviation of 4 min 49 sec.

E. Discussion of the results

As the main quantitative measure of the results the Mean Task Completion Time (MTCT) was chosen, which expresses the average TCT achieved by pairs in each study stage. The results of the performed study reveal a slight difference in MTCT in favor of the new system (Table I). It is difficult to say basing on this results whether or not this difference is statistically significant. However, basing on simple analysis of the obtained times, it can be stated that the direction chosen in the development of the system, consisting of applying AI methods such as fuzzy logic and FCM is proper and should



Fig. 2. Adaptive relations version of the system

TABLE I
RESULTS OF BOTH STUDIES

System	MTCT		σ_{TCT}	
	min	sec	min	sec
1	21	18	3	55
2	19	41	4	49

be continued. Apart from the quantitative results, qualitative results obtained during interviews with participant is also very important at this stage of development. This qualitative results were received from the analysis of recordings, user interviews and questionnaires. At both stages, there were participants that were distanced from such new technological solution, however they recognized the potential advantages of such systems. There are also many voices concerning the decision on implementing such solution on tablets instead of, for instance, one big tactile table. This is actually a discussion beyond this project, because people working on several tablets are more likely to meet in ad-hoc real-life situations than around a huge interactive tabletop. In the current implementation of the system, users generally agreed with the idea of determining relations, and they noticed the intuitive way of displaying it to the user in form of different, more light or dark, shades of one color. The idea of such a solution arose on the basis of preliminary study, where participants were asked to complete the task using pieces of paper. An observation was made then on how people use physical space to organize information. This subject was further investigated and the preliminary ideas for design were published in a workshop paper. This triggered the development of a system based on Fuzzy Inference System and Fuzzy Cognitive Maps. The solution presented in this work is another step towards supporting human cognition with the use of modern technology. The system translates the mutual spatial position of atomic elements of information, which are hints leading to solution in a crime mystery, into their mutual relation, and afterwards allows the user to display these relations, which should enhance the process of making sense of information.

F. Directions for future work

The main direction for future work on this project is an attempt to determine the parameters of the system - both in terms of FIS and FCM. This means further study how exactly should the relation change basing on the input data to give best results. Another area of further development is to find which other information that can be obtained from the analysis of spatiotemporal data in the system should have an impact on the strength of relation. The latter thing mentioned is very intriguing both from theoretical and experimental point of view. It also leads to the last, but not least planned future stage of development of the system. This should be an attempt to generalize the solution to a wide range of problems - so that it can support the sensemaking process with minimal expert knowledge in terms of initial state and parameters of the system.

VI. CONCLUSIONS

In the course of this work, a mechanism supporting collaborative sensemaking with multi-device spatially-aware system was elaborated. This work was preceded by profound research in the field of user interfaces in multi-device environments and was mainly motivated by an emerging need of implementing advanced processing methods in HCI. The results of the user study, compared with the results obtained with the system which did not support adapting the mutual relations basing on user interaction, give a promising insight into the future of such solutions and prove that this step already taken is a step in the right direction. This work needs to and will be further developed.

REFERENCES

- [1] P. Wozniak, N. Goyal, P. Kucharski, L. Lischke, S. Mayer, and M. Fjeld, "RAMPARTS," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, (New York, New York, USA), pp. 2447–2460, ACM Press, 2016.
- [2] M. Fjeld, P. Woźniak, J. Cows, and B. Nardi, "Ad hoc encounters with big data: Engaging citizens in conversations around tabletops," *First Monday*, vol. 20, 2 2015.
- [3] S. Weise, J. Hardy, P. Agarwal, P. Coulton, A. Friday, and M. Chiasson, "Democratizing ubiquitous computing," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, (New York, New York, USA), p. 521, ACM Press, 9 2012.
- [4] M. Bilezikjian, S. R. Klemmer, R. L. Mandryk, J. A. Landay, and L. M. Inkpen, "Exploring a New Interaction Paradigm for Collaborating on Handheld," 8 2002.
- [5] A. F. Blackwell, G. Bailey, I. Budvytis, V. Chen, L. Church, L. Dubuc, D. Edge, M. Linnap, V. Naudziunas, and H. Warrington, "Tangible Interaction in a Mobile Context," *CHI'07 Workshop on Tangible User Int. in Context and Theory*, 2007.
- [6] J. R. Cauchard, M. Löchtefeld, P. Irani, J. Schoening, A. Krüger, M. Fraser, and S. Subramanian, "Visual separation in mobile multi-display environments," in *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, (New York, New York, USA), p. 451, ACM Press, 10 2011.
- [7] P. Hamilton and D. J. Wigdor, "Conductor: enabling and understanding cross-device interaction," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, (New York), pp. 2773–2782, ACM Press, 4 2014.
- [8] A. Lucero, J. Holopainen, and T. Jokela, "Pass-them-around: collaborative use of mobile phones for photo sharing," in *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, (New York), p. 1787, ACM Press, 5 2011.

- [9] A. Lucero, T. Jokela, A. Palin, V. Aaltonen, and J. Nikara, "EasyGroups: binding mobile devices for collaborative interactions," in *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts - CHI EA '12*, (New York), p. 2189, ACM Press, 5 2012.
- [10] J. Cassens, "Cross-Device Interaction," in *AMBIENT 2013 : The Third International Conference on Ambient Computing, Applications, Services and Technologies*, 2013.
- [11] G. Chen and D. Kotz, "A survey of context-aware mobile computing research," 2000.
- [12] D. Reilly, A. Echenique, A. Wu, A. Tang, and W. K. Edwards, "Mapping out Work in a Mixed Reality Project Room," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, (New York, New York, USA), pp. 887–896, ACM Press, 4 2015.
- [13] P. W. Wozniak, B. Schmidt, L. Lischke, Z. Franjic, A. E. Yantaç, and M. Fjeld, "MochaTop," in *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems - CHI EA '14*, (New York, New York, USA), pp. 2329–2334, ACM Press, 4 2014.
- [14] M. Spindler, W. Büschel, C. Winkler, and R. Dachsel, "Tangible displays for the masses: spatial interaction with handheld displays by using consumer depth cameras," *Personal and Ubiquitous Computing*, vol. 18, pp. 1213–1225, 11 2013.
- [15] M. Spindler and R. Dachsel, "Exploring information spaces by using tangible magic lenses in a tabletop environment," *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10*, p. 4771, 2010.
- [16] M. Spindler, "Spatially aware tangible display interaction in a tabletop environment," *Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces - ITS '12*, p. 277, 2012.
- [17] K. Hasan, D. Ahlström, and P. Irani, "Ad-binning: leveraging around device space for storing, browsing and retrieving mobile device content," *Proceedings of CHI 2013*, pp. 899–908, 2013.
- [18] R. Rädle, H.-C. Jetter, M. Schreiner, Z. Lu, H. Reiterer, and Y. Rogers, "Spatially-aware or Spatially-agnostic?," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, (New York, New York, USA), pp. 3913–3922, ACM Press, 4 2015.
- [19] R. Rädle, H.-C. Jetter, N. Marquardt, H. Reiterer, and Y. Rogers, "HuddleLamp: Spatially-Aware Mobile Displays for Ad-hoc Around-the-Table Collaboration," in *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces - ITS '14*, (New York), pp. 45–54, ACM Press, 11 2014.
- [20] P. Kucharski, A. Romanowski, K. Grudzień, and P. Woźniak, "TomSense: Towards Multi-Device Spatial Awareness Based on Independent Plane Sensing," 2016.
- [21] P. Hájek, *Metamathematics of fuzzy logic*. Kluwer, 1998.
- [22] E. Cox, *The fuzzy systems handbook : a practitioner's guide to building and maintaining fuzzy systems*. AP Professional, 1994.
- [23] L. Zadeh, "Fuzzy logic = computing with words," *IEEE Transactions on Fuzzy Systems*, vol. 4, pp. 103–111, 5 1996.
- [24] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets and Systems*, vol. 90, pp. 111–127, 9 1997.
- [25] B. Kosko, "Fuzzy cognitive maps," *International Journal of Man-Machine Studies*, vol. 24, pp. 65–75, 1 1986.
- [26] R. Axelrod, "The cognitive mapping approach to decision making," *Structure of decision*, pp. 221–250, 1976.
- [27] E. I. Papageorgiou and C. D. Stylios, "Fuzzy cognitive maps," *Handbook of Granular Computing*, pp. 755–774, 2008.
- [28] R. Taber, "Knowledge processing with Fuzzy Cognitive Maps," *Expert Systems with Applications*, vol. 2, no. 1, pp. 83–87, 1991.
- [29] M. Styblinski and B. Meyer, "Fuzzy cognitive maps, signal flow graphs, and qualitative circuit analysis," in *IEEE International Conference on Neural Networks*, pp. 549–556, IEEE, 1988.
- [30] J. Craiger, D. Goodman, R. Weiss, and A. Butler, "Modeling organizational behavior with fuzzy cognitive maps," vol. 1, pp. 120–123, 1996.
- [31] J. Aguilar, "A survey about fuzzy cognitive maps papers," *International journal of computational cognition*, vol. 3, no. 2, pp. 27–33, 2005.
- [32] P. Cingolani and J. Alcalá-Fdez, "jFuzzyLogic: a robust and flexible Fuzzy-Logic inference system language implementation," in *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*, pp. 1–8, IEEE, 2012.
- [33] P. Cingolani and J. Alcalá-Fdez, "jFuzzyLogic: a Java Library to Design Fuzzy Logic Control Controllers According to the Standard for Fuzzy Control Programming,"
- [34] K. Hornbæk, "Some whys and hows of experiments in human-computer interaction," *Human-Computer Interaction*, vol. 5, no. 4, pp. 299–373, 2011.
- [35] G. Stanford and B. D. Stanford, *Learning Discussion Skills Through Games*. Citation Press, 1971.

Joint 38th IEEE Software Engineering Workshop (SEW-38) and 5th International Workshop on Cyber-Physical Systems (IWCPS-5)

THE IEEE Software Engineering Workshop (SEW) is the oldest Software Engineering event in the world, dating back to 1969. The workshop was originally run as the NASA Software Engineering Workshop and focused on software engineering issues relevant to NASA and the space industry. After the 25th edition, it became the NASA/IEEE Software Engineering Workshop and expanded its remit to address many more areas of software engineering with emphasis on practical issues, industrial experience and case studies in addition to traditional technical papers. Since its 31st edition, it has been sponsored by IEEE and has continued to broaden its areas of interest.

One such extremely hot new area are Cyber-physical Systems (CPS), which encompass the investigation of approaches related to the development and use of modern software systems interfacing with real world and controlling their surroundings. CPS are physical and engineering systems closely integrated with their typically networked environment. Modern airplanes, automobiles, or medical devices are practically networks of computers. Sensors, robots, and intelligent devices are abundant. Human life depends on them. CPS systems transform how people interact with the physical world just like the Internet transformed how people interact with one another.

The joint workshop aims to bring together all those researchers with an interest in software engineering, both with CPS and broader focus. Traditionally, these workshops attract industrial and government practitioners and academics pursuing the advancement of software engineering principles, techniques and practices. This joint edition will also provide a forum for reporting on past experiences, for describing new and emerging results and approaches, and for exchanging ideas on best practice and future directions.

TOPICS

The workshop aims to bring together all those with an interest in software engineering. Traditionally, the workshop attracts industrial and government practitioners and academics pursuing the advancement of software engineering principles, techniques and practice. The workshop provides a forum for reporting on past experiences, for describing new and emerging results and approaches, and for exchanging ideas on best practice and future directions.

Topics of interest include, but are not limited to:

- Experiments and experience reports

- Software quality assurance and metrics
- Formal methods and formal approaches to software development
- Software engineering processes and process improvement
- Agile and lean methods
- Requirements engineering
- Software architectures
- Design methodologies
- Validation and verification
- Software maintenance, reuse, and legacy systems
- Agent-based software systems
- Self-managing systems
- New approaches to software engineering (e.g., search based software engineering)
- Software engineering issues in cyber-physical systems
- Real-time software engineering
- Safety assurance & certification
- Software security
- Embedded control systems and networks
- Software aspects of the Internet of Things
- Software engineering education, laboratories and pedagogy
- Software engineering for social media

EVENT CHAIRS

- **Bowen, Jonathan**, Museophile Ltd., United Kingdom
- **Hinchey, Mike**(Lead Chair), Lero-the Irish Software Engineering Research Centre, Ireland
- **Szmuc, Tomasz**, AGH University of Science and Technology, Poland
- **Zalewski, Janusz**, Florida Gulf Coast University, United States

PROGRAM COMMITTEE

- **Ait Ameer, Yamine**, IRIT/INPT-ENSEEIH, France
- **Banach, Richard**, University of Manchester, United Kingdom
- **Bensalem, Saddek**, VERIMAG, France
- **Broy, Manfred**, Technische Universitaet Muenchen, Germany
- **Čaplinskas, Albertas**, Vilnius University, Lithuania
- **Carter, John**, University of Guelph, Canada
- **Cicirelli, Franco**, Universita della Calabria, Italy
- **Denney, Ewen**, SGT/NASA Ames, United States

- **Derrick, John**, University of Sheffield
- **Ehrenberger, Wolfgang**, Hochschule Fulda, Germany
- **Eleftherakis, George**, The University of Sheffield International Faculty, CITY College, Greece
- **Fantechi, Alessandro**, DSI - Università di Firenze, Italy
- **Fidge, Colin**, Queensland University of Technology, Australia
- **Forbrig, Peter**, University of Rostock
- **Fortiers, Stephen**, George Washington University
- **Friesel, Anna**, Technical University of Denmark, Denmark
- **Fujita, Masahiro**, University of Tokyo, Japan
- **Golatoski, Frank**, University of Rostock, Germany
- **Gomes, Luis**, Universidade Nova de Lisboa, Portugal
- **Gracanin, Denis**, Virginia Tech, United States
- **Grega, Wojciech**, AGH University of Science and Technology, Poland
- **Gumzej, Roman**, Faculty of Logistics, University of Maribor, Slovenia
- **Havelund, Klaus**, Jet Propulsion Laboratory, California Institute of Technology, United States
- **Hsiao, Michael**, Virginia Tech, United States
- **Kornecki, Andrew J.**, Embry Riddle Aeronautical University, United States
- **Laplante, Phillip A.**, PennState University, United States
- **Letia, Tiberiu**, Technical University of Cluj-Napoca, Romania
- **Li, Jianwen**, Iowa State University, United States
- **Liu, Zhiming**, Southwest University, China
- **Lopez, Oscar Pastor**, Valencia
- **Malloy, Brian**, Clemson University, United States
- **Marwedel, Peter**, Technische Universität Dortmund, Germany
- **Minchev, Zlatogor**, Bulgarian Academy of Sciences, Bulgaria
- **Monostori, László**, Hungarian Academy of Sciences, Hungary
- **Nesi, Paolo**, DSI-DISIT, University of Florence, Italy
- **Obermaisser, Roman**, Universität Siegen, Germany
- **Palanque, Philippe**, ICS-IRIT, University Toulouse 3, France
- **Pu, Geguang**, East China Normal University
- **Pullum, Laura**, Oak Ridge National Laboratory, United States
- **Qin, Shengchao**, Teesside University, United Kingdom
- **Reeves, Steve**, University of Waikato, New Zealand
- **Roman, Dumitru**, SINTEF / University of Oslo, Norway
- **Rouff, Christopher**, Lockheed Martin, United States
- **Rozier, Kristin Yvonne**, NASA Ames Research Center
- **Ryan, Kevin**, Lero-the Irish Software Research Centre, Ireland
- **Rysavy, Ondrej**, Brno University of Technology, Czech Republic
- **Sachenko, Anatoly**, Ternopil National Economic University, Ukraine
- **Sanden, Bo**, Colorado Technical University, United States
- **Seceleanu, Cristina**, Mälardalen University, Västerås, Sweden
- **Sekerinski, Emil**, McMaster University, Canada
- **Selic, Bran**, Simula Research Lab, Norway
- **Sojka, Michal**, Czech Technical University, Czech Republic
- **Sun, Jing**, The University of Auckland, New Zealand
- **Taguchi, Kenji**, AIST, Japan
- **Trybus, Leszek**, Rzeszow University of Technology, Poland
- **van Katwijk, Jan**, Delft University of Technology, The Netherlands
- **Vardanega, Tullio**, University of Padova, Italy
- **Velev, Miroslav**, Aries Design Automation, United States
- **Vilkomir, Sergiy**, East Carolina University, United States
- **Waeselynck, Hélène**, LAAS-CNRS Toulouse, France
- **Zhu, Huibiao**, Software Engineering Institute - East China Normal University
- **Zoebel, Dieter**, University Koblenz-Landau, Germany

The Use of Gamification for Teaching Algorithms

Luiz Ricardo Begosso¹

Douglas Sanches da Cunha²

João Victor Pereira Pinto³

Lucas Teixeira de Lemos⁴

Michel Gargel Nunes⁵

Fundação Educacional do Município de Assis

Centro de Pesquisas em Informática

Assis, São Paulo, Brazil

¹begosso@femanet.com.br;

²professordouglasacunha@gmail.com; ³p-vittor@hotmail.com;

⁴lucas.lemos97@live.com; ⁵michelgargel@hotmail.com

Luiz Carlos Begosso

Fundação Educacional do Município de Assis

Faculdade de Tecnologia de Assis

Assis, São Paulo, Brazil

lbegosso@femanet.com.br

Abstract—This paper presents our experience using gamification principles into the free and open-source learning management system Moodle for aiding and abetting our Computer Science students in learning algorithms. In this work, we used the LMS Moodle and we developed a module with gamification features focused on promoting engagement of students in the learning process of basic concepts of algorithms, data structures and pointers. We conducted a deep study about Moodle and how to implement gamification plugins into the environment. We used and configured HotPotatoes, Games, LevelUp and Badges plugins. We defined the lessons about the specific concepts of algorithms and we created them into Moodle environment. Then we implemented several games, like hangman, crosswords, cryptex, snakes and ladders. We used LevelUp with the objective to gamify the students' learning experience by allowing them to earn experience points to level up in their courses. Badges were used in order to motivate students in their achievements and to show their progress in the courses. In our learning environment, it is possible to have different levels of proficiency in the class and rewards are based on different rules the professor can define; the student can earn more points for some activities or fewer points for other activities. A comparative ranking can be displayed so that students can see their progress, and the professor can look at the log to verify the activities students did and the points they have earned. The results showed us that the use of gamification concepts can contribute significantly to the process of teaching-learning programming concepts to students in the early years, as well for teenager's students without any previous knowledge about programming concepts. This study aims to present the methodology used to carry out our experience and the results obtained with the development and implementation of gamification concepts in a free and open-source learning management system.

Index Terms—gamification; education; computer programming; software algorithms; computer science education

I. INTRODUCTION

THE teaching of algorithms and programming concepts to first-year students has become a critical challenge to En-

gineering and Computer Science courses. It is usual that students face difficulties to understand some concepts they are taught for the first time, such as logical thinking, abstraction, algorithms, data structures, formal computer language, and others. This is potentially quite challenging material that is going to form the basis of the rest of their learning. In few weeks students are introduced to data structures, programming resources, binary trees, sorting, which are examples of very important subjects for them, but many students do not learn those concepts appropriately.

Several efforts have been conducted to get more positive results in the learning process of programming concepts [1]. This paper presents our experience using gamification principles into the free and open-source learning management system Moodle for aiding and abetting our Computer Science students in learning algorithms.

The Millennials, also known as Generation Y, and the post-Millennials, also known as Generation Z, use technology as part of their lives, they are digital natives and have more of a positive view of how technology is affecting their lives than any other generation. Learning is a faster and more flexible endeavor for Millennials and post-Millennials if they can use their smartphone or tablet PC. Both those generations are university students right now.

As described by [2], Millennials and post-Millennials are best suited to modern learning methods and prefer learning in a more relaxed environment, expect instant gratification, and value a fun and flexible learning environment where colleagues are friends. They like to have some control over their development and feel comfortable using technology in the classroom.

This affinity with technology encourages the use of additional tools for supporting the teaching-learning process. In this context, the use of games and simulation environments has taken place in the academia and is getting more and more attention of researchers. The process of using game thinking and mechanics to engage an audience and solve problems is

named “Gamification”. The concept of gamification is associated with the use of game elements in generic contexts. Gamification is the use of game-design elements in a particular task, providing more intense interaction on the exchange of information and encouraging the involvement of the public in a playful way. The concept of gamification has been gaining prominence also in the educational area and this work is situated in this context.

According to [3], some examples of gamification have been used in the area of Information Technology, such as Ribbon Hero, which is an application for the corporate management area, serves to educate users of Microsoft Office 2007 and 2010 how to use the ribbon interface. In another example, the Duolingo, which is a language-learning platform, adopts the experience of accumulating experience points to measure the progress of learning a foreign language by the user.

In this context, gamification techniques that use game design and mechanics can be applied or found in many areas, such as education, corporate environment, entertainment, retail trade, among others.

Tasks that tend to become boring or unnoticed are an object of study for gamification, aiming to become more attractive and provide more intense interactions and experiences by applying gaming techniques in non-gaming environments, engaging the involvement of the public in a playful way. According to [4], the term gamification encompasses the use of game elements in activities that are not strictly a game, that is, the individual thinks and uses game systematics and mechanics, but their action does not determine that he is playing a game.

For [5], gamification takes place from characteristics that we like most in games and incorporate them into our daily activities, so that tasks can be carried out in a fun and exciting way.

Using a set of gaming mechanisms and design techniques in a gamified environment, learning can be encouraged as entertainment because it awakens and increases interest and enhances pleasure while performing a task. It can also increase the content retention and improve motivation for learning.

It is worth mentioning that the techniques of gamification include several characteristics, however, it is not mandatory to apply all of them, since the literature defining the term gamification presents differences of interpretation. According to [6], the most common features found in gamified applications are points, levels, rankings, challenges and missions, medals, achievements, integration, engagement, personalization, feedback, rules and narrative.

Games have great potential to improve the learning experience [7]. For these authors, gamification tends to produce improvements in the understanding, commitment and motivation of users.

In this project, we developed a module with gamification features focused on promoting engagement of students in the learning process of basic concepts of algorithms, data structures and pointers. We used the LMS Moodle, a free and open-

source learning management system, widely used in academia. We conducted a deep study about Moodle and how to implement gamification plugins into the environment. We used and configured HotPotatoes, Games, LevelUp and Badges plugins. We defined the lessons about the specific concepts of algorithms and we created them into Moodle environment. Then we implemented several games, like hangman, crosswords, cryptex, snakes and ladders. We used LevelUp with the objective to gamify the students' learning experience by allowing them to earn experience points to level up in their courses. Badges were used in order to motivate students in their achievements and to show their progress in the courses.

This paper is organized into five sections. Section I is the introduction, while section II describes the learning environment we used to implement the project; Section III describes the details of the implementation of the project; the assessment of the learning environment is outlined in Section IV; and Section V presents our conclusions.

II. THE LEARNING ENVIRONMENT

We decided to use Moodle (Modular Object-Oriented Distance Learning) to apply the gamification concepts because it is a learning management system (LMS) created under the concept of free software that can be installed on different platforms, such as Unix, Linux, Windows and MAC OS. Its development is collaborative by a virtual community, which brings together programmers, designers, administrators, educators and users from all over the world and is available in several languages. The platform has supported Distance Education and face-to-face courses, the formation of study groups, professional training and others [8].

We conducted a deep study about Moodle and how to implement gamification plugins into the environment. We used and configured HotPotatoes, Games, LevelUp and Badges plugins, which will be described below.

HotPotatoes is a plugin created by the Research and Development team at the University of Victoria Humanities Computing and Media Centre, Canada. It contains a package of five tools or authoring programs for the creation of interactive exercises for the Web, named JCloze, JCross, JMatch, JMix and JQuiz; these tools are compiled into one unit, using a sixth application called The Masher. For the implementation of this plugin, it is necessary to understand where the information will be placed (texts, questions, answers, images), since the tools will automatically create the respective webpage for the use of the students.

Games is the second plugin that can be installed in Moodle, in order to provide the creation of educational games. Games plugin is used to simplify the development of gamification concepts in the project. This plugin has several traditional gamification features: hangman, crosswords, cryptex, millionaire, sudoku, snakes and ladders, the hidden picture, book with questions.

For the hangman feature, a keyword is chosen from a glossary or quiz short answer questions and generates a hangman

puzzle. The teacher can set the number of words that each game contains, configure if it shows the first or last letter, or if it shows the question or the answer at the end. Students will need to deduce which word will be explored within the rules of the game, based on the content that was previously studied.

For the crosswords feature, words are taken from either a Glossary or quiz short answer questions and it generates a random crossword puzzle. Teacher can set the maximum number of columns/rows or words that it contains. Students can press the button “Check crossword” to check if the answers are correct. Every crossword is dynamic so it is different for every student.

For the Cryptex feature, it is like the crossword but the answers are hidden inside a random cryptex and the student needs to deduce them, based on the content previously studied.

The Millionaire feature takes words from multiple choice quiz questions and creates a “Who wants to be a Millionaire” style game complete with the three lifelines. Students must answer each question correctly to proceed.

For the Sudoku feature, a sudoku puzzle is presented to the students with not enough numbers to allow it to be solved. For each question the student correctly answers an additional number is slotted into the puzzle to make it easier to solve.

The hidden picture feature randomly grabs an image from a glossary and hides it behind panels. When the student answers a question correctly, a portion of the image is revealed.

For the Snakes and Ladders feature, students have to traverse a traditional “Snakes and Ladders” board by answering questions taken from either a Glossary or quiz short answer questions. As they get an answer right, the dice are rolled and a random number displayed. The game piece is moved ahead of that many squares. If the game piece is in the bottom of a ladder and the answer is correct, it goes to the top. If the game piece is in the head of the snake and the answer is wrong, it goes back to the tail.

Book with questions feature controls the progression of the student, he can go to the next chapter only if he answers the questions correctly.

We also used the LevelUp plugin, with the objective to gamify the students’ learning experience by enabling learners to gain experience points for participating in their courses, increasing engagement and participation by motivating students to progress towards the next level and rewarding their efforts by congratulating them for reaching the next level. LevelUp plugin allows the teacher to use the leaderboard to leverage competitiveness while keeping it friendly and motivating, unlock access to course content when a certain level is reached, and substitute experience points for other images to make the learning process even more attractive for the student.

We used the Badges plugin to allow the teacher to show student’s progress awarding him with badges, based on a variety of chosen criteria.

III. PROJECT IMPLEMENTATION

Our institution enrolls approximately 2,200 students in 11 programs. STEM-related degree programs are the most traditional and we have around 330 students enrolled at Computer Science course, with a typical first-year class of around 100 students. In the first year, the Department of Computer Science offers an introductory programming course: Algorithms I, which is the expected starting point for students majoring in our Department. It is a full-year typical introductory course with no previous programming experience required. In the second year, we offer the course Algorithms II, closing the fundamentals for logic and programming basics.

From our previous experience, we know that it is difficult for students to understand some concepts they are taught for the first time, such as logical thinking, abstraction, algorithms, data structures, formal computer language, and others. This is potentially quite challenging material that is going to form the basis of the rest of their learning. In few weeks students are introduced to data structures, programming resources, binary trees, sorting, which are examples of very important subjects for them, but many students do not learn those concepts appropriately.

Our project is divided into five phases, as described in Figure 1.

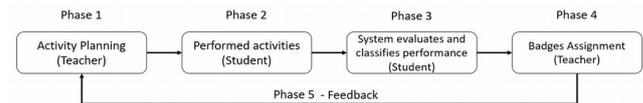


Figure 1 –Phases of the Project

During Phase 1, teacher defines the learning outcomes that students should be able to know in order to complete a study stage successfully; the teacher must decide which concepts and skills students should have at the end of a given learning period.

When Phase 2 begins, students access the learning environment and perform the activities proposed by the teacher.

On Phase 3, the system evaluates and classifies students’ performance, based on the activities performed by each student and on the results achieved by each student.

During Phase 4, the teacher assigns a badge to the students, according to their performance and experience points gained.

For the feedback phase, the teacher can check what aspects students presented most difficult, based on reports generated by the learning environment about students’ performance, and make a new plan of activities, returning to Phase 1.

In this project, we decided to apply games concepts in the Moodle free software platform, to verify if our students would have more success in learning those important concepts. So, we implemented this course covering the following subjects: Declaration and Manipulation of Variables, Repetition Structures, Data Structures, Pointers, and List Structures. We integrated the HotPotatoes, Games, LevelUp and Badges plugins into Moodle platform, using textual and videos resources fo-

cusing on the practical concepts of the contents, and the games were used for engaging students in the learning process.

We implemented four packages from HotPotatoes: JCloze, JCross, JMatch and JQuiz. The JCloze package creates gap-fill exercises, for example, student need to complete sentences about some taught concept. Unlimited correct answers can be specified for each gap, and the student can ask for a hint and see a letter of the correct answer. The JCross package creates crossword puzzles which can be completed online, with words from theory taught; the teacher can configure grids of any size and the system places the words in columns automatically. The JMatch package creates matching exercises, where students can match vocabulary to pictures or translations, or ordering sentences to form a sequence or a conversation. The JQuiz package creates multiple-choice and short-answer quizzes; specific feedback can be provided both for right answers and predicted wrong answers.

For the LevelUp plugin, there are general and specific configurations the instructor needs to do. Initial settings include General, Ranking, Cheat Guard, and Logging settings, and there are different tabs, like Ladder, Report, Log, Level, Rules, Visual and Settings. At Settings tab, teacher can enable the student to gain Experience Points and reach new levels. In the configuration of the Student Ranking, instructor controls whether participants can see each other's name and avatar. Neighbors are the participants ranked above and below the current user. For instance, when choosing 'Display two neighbors', only the two participants ranked directly higher and lower than the current user will be displayed. The rank is the absolute position of the current user in the ladder. The relative rank is the difference in experience points between the user and their neighbors.

At Cheat Guard setting, teacher can enable it to limit the maximum number of actions that will count for Experience Points during the time frame given to the student. Any subsequent action is ignored. Teacher can configure the minimum time required between identical actions (an action is considered identical if it was placed in the same context and object; for example, reading a forum post will be considered identical if the same post is read again).

At Visual tab, teacher can configure images that are used as badges, which are assigned to each student, according to the level they reach the experience points, showing their specific performance. Figure 2 presents the badge images we configured for each level.

At Infos tab, teacher can see all the values used to identify each level and its respective image used to customize the badges. This tab records all the actions performed by students, so the teacher can create motivational phrases, which are attributed to each student when he reaches a new level and receives a new badge.

Finally, at Rules tab, teacher defines the rules to compose all the activities that will be evaluated by the LevelUP plugin. In this tab, the teacher can configure several rules considering all the activities students have to do, setting values for experi-

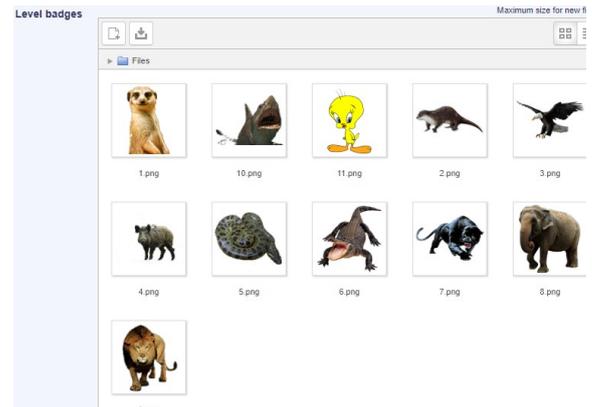


Figure 2 –Badge images used for each level

ence points and identifying which events are triggered as students perform actions in the course. Once each rule is included by the teacher, he can add or remove activities and resources which LevelUP plugin will record as an action to be taken on the student's experience point.

When students access their dashboard for the first time, on the left side they have information about grades and level reached, on the right side they have access to the lessons, exercises (traditional and games) and homework. Figure 3 presents the student view of the learning environment.



Figure 3 –Student view of the learning environment

On the dashboard of the course, students have access to the concepts taught, for example, Figure 3 illustrates the concepts about declaration and manipulation of variables. Below the first concepts, students have questions to answer and the first game: crosswords. On the left side of the dashboard, students have their level, which changes to different images as they participate in the course and gain experience points to level up.

Instructors and teachers have access to students' progress dashboard and they can verify the reports about students' progress, students' log of use, change the rules for reaching different levels and change general settings. Figure 4 illustrates how the instructors can see students' report, with students' name, the level reached, experience points gained and progress. All the environment is configured by the instructor, so we can check which level the student reached, how many

points he has earned until that moment, and the student’s progress compared to other students.

First name / Surname	Level	Experience points	Progress
[Redacted]	10	1190	1100/1190
[Redacted]	10	1354	1354/1354
[Redacted]	10	2150	2100/2150
[Redacted]	10	1301	1301/1301
[Redacted]	9	1071	108/121
[Redacted]	8	917	79/121
[Redacted]	7	747	20/121
[Redacted]	7	821	100/121
[Redacted]	7	749	20/121
[Redacted]	6	670	70/121
[Redacted]	6	695	65/121

Figure 4 –Students’ progress dashboard

Figure 5 illustrates some of the games students need to participate for gaining experience points, the first game is the crosswords, the second game is the hangman, the third game is a quiz and the fourth game is the cryptex.

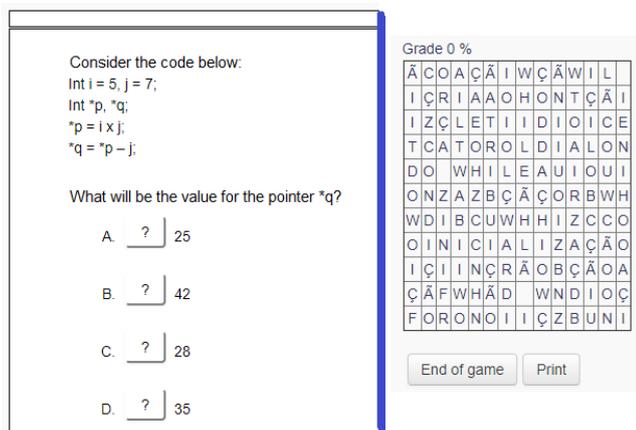


Figure 5 –Examples of games implemented in the project

For each game, students have a specific number of tries, configured by the teacher, and successful results take students to the next level.

IV. ASSESSMENT OF THE ENVIRONMENT

In order to know better the effectiveness of our gamification environment for promoting engagement of students in the learning process of basic concepts of algorithms, we applied it for our first-year students in the Computer Science course. The study was conducted with two groups, each one of them with 22 students (n = 22), of both sexes. The first group was called by “Game Group” and the second group was called by “Test Group”.

The Game Group explored the gamified learning environment during two weeks, and then they performed the activities planned in the environment. At the end of the experience, students of this group were tested for the concepts discussed in the environment.

At the same time, the Test Group was being taught the same concepts by professors using the traditional methods. At the end of this stage, students were tested for the same concepts of the first group.

The analysis of the data was performed with a paired t-test on the same sample unit and the objective was to verify if there was a significant difference between the two groups. Table I presents the results each group reached on this test.

Table I - t-Test: Paired Two Sample for Means (p < 0,05)

	Number of Individuals	Mean Score	Standard Deviation	P-value
Game Group	22	8.0	2.07	0.0015
Test Group	22	5.5	3.26	

We can see in Table I that the Game Group obtained a mean score significantly higher than the Test Group (8.0 ± 2.07 vs 5.5 ± 3.26 , $p < 0.05$). These results were considered very positive because the gamified learning environment increased the content retention and improved motivation for learning.

After having these encouraging results, we applied new questionnaires for students to verify their perception about the qualitative results for the learning environment.

Students explored the content of the modules and, at the end of each stage, they were asked to respond to a questionnaire to evaluate several aspects of the environment. Figure 6 illustrates what students thought about the usability of the tool.

We can see in Figure 6 that 35.3% of the students answered that the tool is very easy to use; 23.5% of the students answered that the tool is easy to use; 23.5% of the students answered that the tool is moderately easy to use. At the other end, only 17.6% of the respondents answered that the tool was reasonably moderate-difficult to use.

Then, we asked the students if the tool can help in the learning process of basic concepts of algorithms. Figure 7 illustrates their answer about this question and we can see that 76.5% of the students answered that the tool has a very great

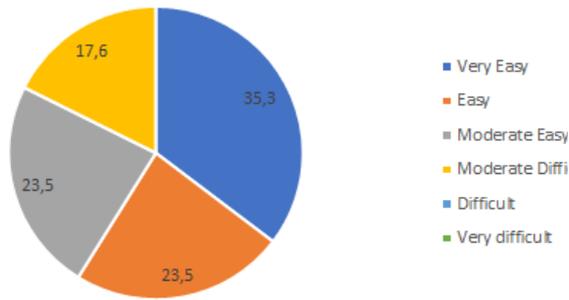


Figure 6 –Degree of difficult in using the tool

contribution to the learning of basic concepts of Programming; 17.6% of the students answered that the tool has a great contribution to the learning of basic concepts of Programming. At the other end, but not least important, only 5.9% answered that the tool has a moderate contribution for their learning process.

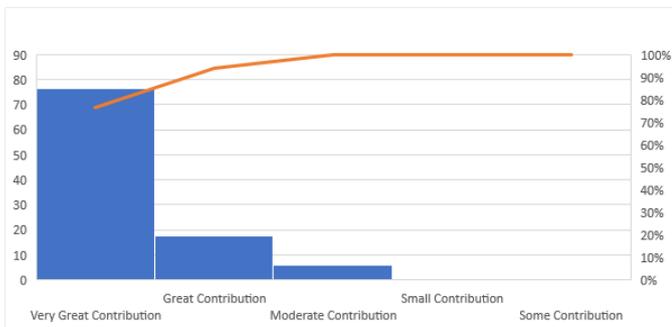


Figure 7 –Contribution of the tool for the learning process

We also asked students to answer which approach they prefer better to learn, traditional or traditional with gamification tools. Figure 8 illustrates their answer about this question and we can see that 82.4% of students prefer to learn by a traditional approach with motivational tools using games; only 11.8% answered that they learn best with traditional classes and 5.8% answered that they prefer learning with other methods.

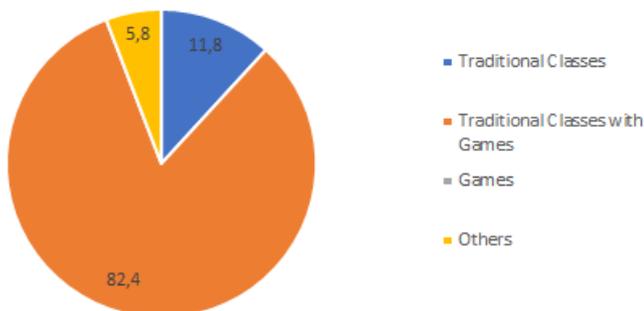


Figure 8 – Students' preference for learning approach

V. CONCLUSIONS

This work aimed to present our experience in using gamification features in the LMS Moodle, focused on promoting engagement of students in the learning process of basic concepts of algorithms, data structures and pointers. We conducted a deep study about Moodle and how to implement gamification plugins into the environment. We used and configured HotPotatoes, Games, LevelUp and Badges plugins. We defined the lessons about the specific concepts of algorithms and we created them into Moodle environment. Then we implemented several games, like hangman, crosswords, cryptex, snakes and ladders, hidden picture, book with questions and others. We used LevelUp with the objective to gamify the students' learning experience by allowing them to earn experience points to level up in their courses. Badges were used in order to motivate students in their achievements and to show their progress.

In order to assess our environment, we conducted an experience with two groups of students, the first one used the gamified learning environment and the second one was taught the same concepts by professors using the traditional methods. The analysis of our results showed that the group that learned the algorithms and programming concepts using our gamified learning environment had expressive higher grades than the second group, what was considered very positive.

Also, for our qualitative assessment, students reported important results, as they liked the usability of the environment and most of them answered that they prefer to learn by a traditional approach integrated with motivational tools using games.

Our gamified learning environment engaged students through tools that easily facilitate social learning and knowledge sharing through forums, chat, blogs and games, increasing students' content retention and improving their motivation for learning, resulting in higher final learning results, comparing with students taught with traditional methods.

The experience of working with free and open-source software distributed under the GNU license was very positive because we concluded that Moodle is more than just an application for distance learning solutions, it is a widely used tool that meets several needs of gamification concepts and any new implementation through the use of plugins for new solutions.

The use of LevelUp was crucial because it caused a competition with students, creating a social pressure for increasing student's level of engagement. The use of badges illustrated the visual representations of merits and achievements, providing feedback to students on their performance as they earned points within the environment, bringing feelings of competence.

The use of Badges was important because it motivated students in their achievements and created a comparison of their progress in the courses.

The results gathered on this project are very important and positive. They can serve as a basis for the academic commu-

nity to start the development of more gamification environments for teaching computer programming, making the learning process for first year's students less difficult.

Based on our study, we conclude that the use of gamification has an important role in the teaching of algorithms and programming concepts to first-year students of Engineering and Computer Science courses, as it is more attractive to youth students and provides more intense interactions and experiences than traditional methodologies. Our gamified teaching environment increased students' interest and enhanced pleasure while students needed to perform a task, increasing the content retention and improving motivation for learning.

REFERENCES

- [1] Begosso, L. R., Begosso, L. C., Begosso, R. H. "An approach for the use of Learning Objects in teaching Computer Programming concepts". In 2016 IEEE Frontiers in Education Conference, Erie, PA, pp. 1-8, 2016. <https://doi.org/10.1109/FIE.2016.7757619>.
- [2] Werth, E. P., Werth, L. "Effective Training for Millennial Students". *Adult Learning*, 22(3), pp. 12-19, 2011. <https://doi.org/10.1177/104515951102200302>.
- [3] Uskov, A., Sekar, B. "Serious games, gamification and game engines to support framework activities in engineering: Case studies, analysis, classifications and outcomes". In IEEE International Conference on Electro/Information Technology, Milwaukee, WI, pp. 618-623, 2014. <https://doi.org/10.1109/EIT.2014.6871836>.
- [4] Nah, F. FH., Zeng, Q. , Telaprolu, V. R., Ayyappa, A. P., Eschenbrenner, B. "Gamification of Education: A Review of Literature". In *HCI in Business*, vol 8527, Nah F.FH. (eds), Lecture Notes in Computer Science: Springer, pp.401-409, 2014. https://doi.org/10.1007/978-3-319-07293-7_39.
- [5] Freitas, S. A. A., Lacerda, A. R. T., Calado, P. M. R. O. , Lima, T. S., Canedo, E. D. "Gamification in Education: A methodology to identify student's profile". In 2017 IEEE Frontiers in Education Conference, Indianapolis, IN, pp. 1-8, 2017. <https://doi.org/10.1109/FIE.2017.8190499>.
- [6] Surendeleg, G., Murwa, V. Yun, H., Kim, Y. S. "The role of gamification in education – a literature review". In *Contemporary Engineering Sciences*, Vol. 7 (29), 1609-1616, 2014. <https://doi.org/10.12988/ces.2014.411217>.
- [7] Barata, G., Gama, S., Jorge, J., Goncalves, D. "Engaging engineering students with gamification". In *Games and virtual worlds for serious applications (vs-games)*, 2013 5th international conference, p. 1-8, 2013. <https://doi.org/10.1109/VIS-GAMES.2013.6624228>.
- [8] Moodle. Documentation about Moodle Platform. Retrieved from <https://moodle.org>. 2018.

Automated generator for complex and realistic test data—a case study

Richard Lipka

NTIS - New Technologies for Information Society
Faculty of Applied Sciences, University of West Bohemia
Plzen, Czech republic
Email: lipka@kiv.zcu.cz

Tomas Potuzak

Department of Computer Science and Engineering
Faculty of Applied Sciences, University of West Bohemia
Plzen, Czech republic
Email: tpotuzak@kiv.zcu.cz

Abstract—Some type of tests, especially stress tests and functional tests, require a large amount of realistic test data. In this paper, we propose a tool JOP (Java Object Populator) that uses a pseudorandom number generator in order to create test sets of complex Java objects, that can be automatically generated and directly used. Along with that, we also show usage of this tool in case study focused on performance evaluation of a real cashier system.

The tool is designed to be able to set simple attributes of any Java object and in many cases also to create complex structures when objects are connected via references. Random values are created using the rules that are added to the class definition in form of annotation to each attribute. Using this tool simplifies creating of tests, as the tester does not need a detailed knowledge of data structures. The specification of expected values is delegated to the designer of the data model and becomes the part of the model. Furthermore, as the data objects are created at runtime, using reflection, the tests do not have to be changed when data carrying objects are modified.

I. INTRODUCTION

IN ORDER to ensure the reliability of each application, testing is a vital part of the software development process. As applications are becoming more complex — especially in terms of using many different existing libraries — and there is a pressure for fast development, one of the most important issues is fast automation of each part of software development process. Many kinds of testing, such as unit tests, are performed automatically, without the need for a human tester. However, the creation of the tests is still mostly a manual process, where a lot of code has to be created.

One of the issues of the testing is obtaining the test data. In some types of tests, for example, when the user interface is being tested, only simple values like numbers, dates, or strings are used. However, when unit tests are focused on the core of the application, it is often necessary to work with the creation of non-trivial testing objects, that are composed not only of simple attributes mentioned before but also contains references to other objects and creates arbitrary complex structures. This is even more significant in stress testing and benchmarking, when a large number of the test objects is required, to avoid biases caused by caching of too similar data. In such cases, testers have to manually prepare all instances, before the tests can be performed.

Furthermore, creating a large set of data is a tiring and repetitive task, so testers help themselves by using random generators that are provided by most programming languages. These generators can be easily used for creating test data, but usually, are designed only to generate numbers. In order to use them, test programmers are bound to have a detailed knowledge of data structure to choose appropriate parameters of the generators. Some attributes can be dependent on the others (for example weight and size of the object). Consequently, the tester has to understand these dependencies. Furthermore, the test created this way contains a lot of code strongly dependent on the structure of data carrying objects. When the implementation of these objects is changed, parts of the tests that set up data have to be revised and changed accordingly. This is reducing benefits of automated testing and forces the testers to return to the tests with each new version.

Another problem is that the characteristics of the data for the test are usually not written anywhere in an explicit form, only as the parameters of the random number generators. This makes updates of the tests more difficult, as the tester has all the time understand both structure of the domain objects and character of the test data.

In order to address described issues, in this paper, we propose a tool that enables to generate test sets of complex Java objects using pseudorandom numbers generators and annotations in source code of the applications under tests describing the possible values of the attributes.

II. ISSUES OF TEST DATA GENERATING

A. Simple motivational example

Consider an application dealing with receipts. Each receipt is represented by an instance of class `Receipt`, with several attributes, such as `date`, `total price`, `salesman` and also a list of items that are on a receipt — each represented again by an instance of class (in this case the `item`). When we want to perform a stress test that adds, removes, or look up for receipts in the database, we need to generate several hundreds or thousands instances of these classes. In the same time, the application might perform other tasks, such as sending data over the network or calculating aggregated values from the receipts. In order to investigate whether the application

behaves correctly, the provided instances should be as realistic as possible.

The classical approach would be to create a generator of receipts and items that will ensure that all attributes of each receipt are set up properly. In order to do so, the creator of the test has to know all attributes that should be set, and also to have a knowledge of their characteristics. Often, the attributes are not a primitive data types but references on other objects, and if the generator should serve its purpose, it should handle this as well. Such generator can be used as long as the class `Receipt` is not changed. In the typical application, there can be dozens of domain classes like the `Receipt`, that serves mainly to carry data and that need to be generated during testing. Thus creating a generator to all of them usually is a time-consuming work.

We would like to have a tool that will be able to create instances of the `Receipt` class in one invocation, according to the characteristics provided in a human readable and understandable form. When random number generators are used in the code of the test, it is not obvious what the meaning of their parameters is without analysing the methods of the generator. It would be useful to have a declarative way of specifying these parameters in one place.

B. Test Data Sources

Data used for testing can be obtained from several sources, and their selection depends strongly on the purpose of testing. For example in unit tests, the most common way is to choose data in the way that the extreme values or decision points of the methods under testing are explored. On the other hand, for the purpose of integration tests or stress tests, it is important to use as realistic data as possible, in order to mimic the real usage of the tested software. We would like to have a tool that will allow generating a large number of instances of domain objects, with as little work of the tester as possible.

The realistic testing data can be obtained in three basic ways. They can be prepared in advance manually, they can be obtained from the production application, or they can be randomly generated.

Manual preparation of the test data is often necessary, but it is a tedious and error-prone work, so testers are usually looking for some way of automation. One option is to automatically capture the data from existing application and reuse them during the testing. This is relatively simple when the application under test and the application used for obtaining the data are the same, otherwise, a conversion is necessary. But either way, if the data are prepared manually or captured from the application, they are a static set that cannot be easily adapted to the changes in the application under test and that cannot be scaled according to the need of testers.

The randomly generated data have a great advantage in scaling, as when the generators are ready, it is easy to create an arbitrary amount of data. They can be also easily parametrized. Furthermore, they might be configured to create different data sets in each run or can create exactly the same sequence of data without the need to store a lot of testing objects.

III. TOOL DESIGN

Our main goal is to create a tool that will allow generating testing data from the definition of classes, enriched by the specific annotations. This way, the structure of data is easily visible from the application source text and is not hidden in the separate source code of tests. The same definitions are used in all tests in order to help the tester to avoid errors caused by code repetition and corrections only in some copies. It can also help to avoid the need to rewrite tests when domain classes are changed. As long as test requires valid instances of domain classes, the tool can provide them.

From the testing point of view, the new annotations serve as a detailed specification of the data type of each generated attribute. For example, instead of working only with information that an attribute is a `double` number, the annotations can specify that the value has for example a normal distribution with specified mean and standard deviation, and when necessary also with a minimal and maximal value that crops the highest and lowest values (in cases like human height). Furthermore, it might be specified that the value is a function of another attribute. The annotations can determine the desired structure of strings or, in more general cases, the characteristics of string language. In case that the attribute is a reference to another object, it is possible to specify the instance or the class that should be used to generate an instance that should be used in the reference, instead of filling the reference with `null` value as is common in contemporary mocking tools.

A. Structure of the Tool

The tool is composed of three main parts — the class analyser, the testing API and the random generators.

The class analyser is responsible for loading the classes that should be instantiated for the testing and searching for their published attributes (the attributes that have corresponding public getters and setters) and the annotations connected to them. The analyser also processes the dependencies between the classes, in order to be able to generate references to other objects. As the annotations are compiled into the bytecode, they can be directly accessed by reflection mechanism and there is no need for direct analysis of the source code of the classes.

The testing API allows testers to easily generate large collections of random data and access the annotations that specify the behaviour of the generators. This is achieved through objects populator. The populator provides a method `populate()` that takes a descriptor of a required class as a parameter and returns the desired number of automatically generated instances in a collection.

The last part, the random generators is responsible for generating primitive datatypes and instances of referred classes. Primitive data are created according to the rules given in annotations and instances of referred classes are created according to selected strategies. It is possible to generate new instances, search among already generated instances or just fill the reference with `null` value.

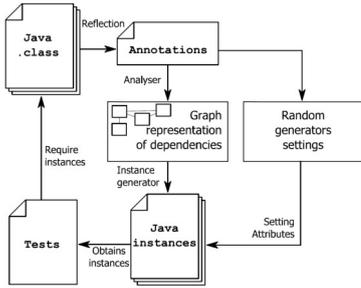


Fig. 1. Basic usage of the JOP.

B. Using the JOP Tool

Our tool is intended to be used as a tool that simplifies writing of stress tests, benchmarks, and unit tests. So far it is designed to work with JavaBeans, i.e. classes that have setters and getters for each attribute and are intended for carrying data within the application. In order to work, the tool also requires an existence of public constructor without parameters. Because the generating of the instances is automated and depends on the reflection, it is necessary to set up attributes of each instance using the setters, as each setter deals only with one attribute. In multiple parameter constructors, it is not possible to find pairing between parameters and attributes and thus to choose the order of generated values that would be used in the constructor.

The process of JOP usage is shown in Fig 1. The tool can work with general JavaBeans, but without adding additional annotations (described in further sections), it cannot create data that will resemble the real ones – only to generate data from the whole space of each attribute type. The first thing the testers or the programmers have to do is to create appropriate annotation (see Section IV) of each attribute they want to randomly generate. Then, when the tests are created, testers can use generating methods and obtain the list of the instances that can be used for further testing. If no annotation is provided, the tool will behave differently for references and for primitive datatypes. References are by default set on `null`, primitive datatypes are generated with uniform distribution on the whole interval of the type.

IV. CLASS ANALYZER

The class analyser is responsible for two main tasks — reading annotations in order to set up the generators for the primitive attributes and analysing the structure of the generated classes in order to determine the instances generating order, setting up references, and generate dependent attributes. While parsing annotations is a simple task, working with dependencies brings several problems.

A. Types of Dependencies

There are two types of dependencies that may influence the process of generating data. The most straightforward is a dependency of one attribute on one or more other attributes from the same class. This can happen even when generating

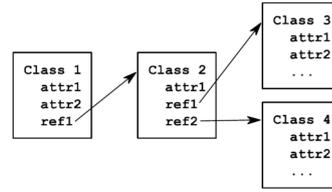


Fig. 2. Visibility of the attributes.

is not recursive and all references are only set to `null`. The more complicated situation is in the case of dependency on another class or on an attribute from another class. This can happen only when the generating of the data is recursive and the referenced classes are generated as well.

In our current implementation, the attribute can be dependent only on the attributes of directly referred classes, not on arbitrary method calls from the referred classes. We have decided to use this limitation in order to be able to create the order of generating only from the attribute definitions, without the need for analysis of the source text of all methods. An example is on the Fig. 2. In annotations within Class 1 it is possible to use `attr1` and `attr2` from Class 1 and `attr1` from Class 2. Attributes from classes 3 and 4 are not accessible.

B. Dependencies Within One Class

In the case of dependency within one class, the ordering of generating operations is straightforward. The analyser is generating values in the following order:

- 1) Create a set of all attributes A and an empty set N .
- 2) Find all primitive attributes without any dependencies and generate their value. Remove all such attributes from the set A and add them to the set N .
- 3) Search the set A for attributes dependent only on the elements from the set N . Move each such attribute from the set A to the set N and in the same time, generate its value (because it depends only on attributes from the set N , all required values have to already exist).
- 4) Repeat step 3 as long as the attributes from the set A are being moved.

If no attribute was moved and the set A is not empty, the remaining attributes contain circular dependency and cannot be resolved. In this case, the generating process throws an exception with the message which attributes and annotations caused the problem. It is important to note here that the mentioned dependency is caused by a tester, when setting up annotations — by claiming that attribute X should be generated as a function of attribute Y and, in the same time, that attribute Y should be generated as a function of attribute X . Removing this dependency does not require changes in domain classes, only in test data definition.

C. Annotations for the Class Dependency

It is possible to use the annotations in order to control generating of the whole graph of dependent objects. We are

using three strategies to do so.

At first, it is possible to forbid the generating of value for selected attribute, using `@Ignore` annotation. In such case, attribute will be skipped and no value will be assigned to it (if it contains a default value, it will not be changed).

It is also possible to set the value of the attribute to `null`, which is a common strategy of many data generating tools. If attribute contains different reference, it will be replaced by `null` reference. If this annotation is used for numerical type, value 0 will be used instead of `null`.

When the reference on new instance of the class should be created and assigned to the reference, annotation `@NewInstance` can be used. When the class contains a default constructor, it can be used directly. When the class contains multiple constructors or a factory method, it can be marked with `@Constructor` annotation, specifying which constructor will be used for instance generating. Parameters of such constructor or factory method can be annotated in the same way as attributes of the class, so the tool can generate their values. If no annotation is specified, the default values for each type will be used.

This annotation has to be paired with another one, that specifies the class provider. The provider is responsible for creating or obtaining desired object. There are several types of providers:

- `@TargetClass` specifies the name of the class which will be used to create new instance.
- `@RandomClass` specifies the list of classes and the probability of each one. This allows to select randomly one of several implementations.
- `@CustomClassProvider` specifies the class that is implementation of `@ClassProvider` interface and serves as a factory for creating of the instances. This annotation serves for using manually created data generators and adding them to the data generating process. The tester can create his own implementation of the generator in case that no approach provided by our tool is suitable for his or her needs.

```
@NewInstance
public Student student; // new instance of
    Student class using default constructor

@NewInstance
@RandomClass(value={Student.class,
    Teacher.class}, probabilities={0.75,
    0.25})
public Person person; // new instance of
    Student or Teacher class

@NewInstance
@CustomClassProvider(RandomStudentProvider.class)
public Student student; // new instance using
    RandomStudentProvider factory
```

Instead of creating always new instances, it is also possible to assign one of the created instances. In such case, annotation `@SearchInstance` can be used.

During the data generating, all created objects (when annotation `@NewInstance` was used) are stored, so it is possible to search among them and use them repeatedly. `@SearchInstance` allows to specify criteria for searching among the existing objects and assign the reference to the annotated attribute. When this annotation is used alone, it will search first suitable instance (the instance of the appropriate class) and assign the reference to it. It can be also combined with tester specified annotation that will specify the rules for selecting required instance.

As there is no simple and general way how to create an annotation for instance search, we have decided to delegate this work to the matcher class that the tester will have to implement. In the tool, simple `InstanceMatcher` is prepared and the tester can use it to implement his own class that is able to decide, which searched instance is suitable. For each implementation, the corresponding annotation is automatically created and can be used immediately.

```
@SearchInstance
public Student student; // first existing
    instance of Student class will be used

@SearchInstance
@RandomStudentClass(age = 26)
public Student student; // user annotation,
    first instance of Student class with
    appropriate age will be used
```

In this example, tester implemented `RandomStudentClass` matcher that can compare the age attribute of provided `Student` instance and determine if the instance fulfills the criterion.

D. Generating the Dependent Instances

When the recursive generation is used, the whole process is divided into two steps. In the first step, all instances are generated and connected via references. In the second step, the dependencies between their primitive attributes are resolved and their values are generated. It is possible that the data objects contain a circular dependency, but, unlike the circular dependency between attributes, this can be solved by using strategy for searching instances.

The algorithm for generating instances works as follows:

- 1) Start from the generated class (the class that was required for the test).
- 2) Load the class annotation.
- 3) Check if the class is already in the dependency graph. If it is so, mark this dependency as `null`. If not so, add class as a node in the dependency graph. Store all annotations in the node.
- 4) Search for all attributes with the `@NewInstance` strategy. Process each such class recursively from the point 2.

This creates a tree of dependencies, with all classes that should be newly generated. Note, that the `@NewInstance` annotation means, that the new classes are always created.

Because of that, circular dependency is not allowed with this annotation, since it would lead to an infinite recursion. Instead, such references should be set to `null` or filled with instances that already exist.

When the dependency graph is finished, the creation of instances of all classes starts. The analyser keeps collections of references for each generated class that will be used for the `SearchInstance` annotation. The generating is performed using the dependency graph in following way:

- 1) Create an empty set A for generated attributes and empty set S for attributes with `@SearchInstance` annotation for each analyzed class, the queue of already created instances Q , set of created instances I and dependency graph G .
- 2) A shared copy of the collection of existing instances I_G is created.
- 3) Start from the class that should be generated, the instance of this class is added to the set I and to the queue Q .
- 4) First instance from Q is taken.
- 5) All attributes which should be generated are stored in the set A .
- 6) All attributes which should be searched (`@SearchInstance`) are stored in set S .
- 7) All attributes with `@NewInstance` annotation are checked if they are not causing circular dependency in graph G . If they cause circular dependency, their value is set to `null`. In the opposite case, the new instance is created and inserted into the queue Q and set I . The class is also added as a new node to the dependency graph G .
- 8) When the queue Q is empty, the algorithm ends. Otherwise it continues from step 4.

Because the searching of instances can be performed according to specified criteria, it is necessary to generate the values of the attributes, which are not dependent on the others. In order to find them, an empty set N is created. Then, all attributes from each version of A set for each class in I are analyzed. When the attribute has no dependency, the values are generated. These attributes are then moved from the set A to the set N .

Now it is possible to search instances in the sets I and I_G for each attribute from the set S . If no suitable instance for some of the attributes is found, its value is set to `null`. Otherwise, the reference on the instance is set to the attribute.

Finally, it is possible to generate the values for remaining attributes from the set A . To do so, it is necessary to go through the attributes in A_i sets for each instance in the I set and generate the value of the attributes using the algorithm described in section IV.B.

V. PRIMITIVE VALUE GENERATORS AND ANNOTATIONS

We can divide generators into three main groups:

- 1) Number generators are responsible for generating any numeric value, integer or real.

- 2) Text generators are used for creating strings, according to the rules based on the length, language or structure of the string.
- 3) Object generators are responsible for generating of Java objects with the specific structure, such as `Color`, enumerate types or logic values.

Each attribute can be annotated with one specification of its values. When no annotation is used, the attribute is ignored (as if `@Ignore` annotation is used). The annotations for attribute generation can be combined with annotation of populators, which can be used to specify how the generated value will be used in the attribute. The populator annotations are used mainly for populating arrays and collections with primitive type values.

A. Number Generators

Java has a capability for generating random numbers, however it contains only a limited number of generators for different probabilistic distributions. We are using *Uncommons Maths* [1] library that is available under Apache Software Licence v. 2.0 and that provides among others a set of random number generators. As each distribution requires different parameters, special annotation (and corresponding generator) is created. Currently, the tool supports 8 different parametrized distributions.

For assigning a number generator to an attribute, appropriate generator annotation is used. Both continuous and discrete generators can be used for each data type.

As was mentioned before, the numerical attributes can be dependent on each other. In such case, annotation `@Expression` is used, in combination with other annotations for value generating. Annotation contains an expression that is evaluated when other attributes are generated.

```
@Expression (" rnd1 * atr1 + ref1.atr2 ")
protected int rnd1 ;
```

When evaluated, the value of `rnd1` and `atr1` will be searched in the instance where this expression is evaluated and the value of the `atr2` will be searched in the instance referred in the `ref1` attribute. The value of the attribute will be then determined as result of the provided expression. The expression can process basic arithmetic, as well as invocation of functions from `Math` library.

B. Text Generators

Generating realistic strings is a common problem, solved for example by `RandomStringUtils` class from Apache Commons project [2]. However, this class can only generate a random string of given length, with the discrete uniform distribution of probability of each character occurrence. It is possible to choose which characters will be used and which omitted, but there are no other setting possibilities. Such strings are not very realistic representation of words in any language and it is difficult to use them as a representation of other string entities (such as colour codes) as well. We are using two types of Markov chain generators:

1) *Language based*: Language based string generators use Markov chains with given corpus. The corpus serves to determine the probability of one letter following another letter or sequence of letters [3]. We have corpuses for English and Czech languages, but it is possible to use other corpuses provided as files.

2) *Pattern based*: Using arbitrary table itself allows to create Markov generators based on patterns, but it is not the most convenient way of doing so. Thus we have created a generator based on regular expressions. This generator is working in the similar way as Markov based, but instead of using probability table, it transforms the regular expression to Finite State Machine. The transitions of FSM represent the generating of next character to the string, but in this time, the probability of all characters in each state is equal.

C. Populators

Populators serves to simplify the generator logic and to allow to use each generator for any data type or data collection. Due to this, generator does not need to know anything about the type of attribute they are generating for and delegates this work on populator, which has the full knowledge of attribute declaration and can set data to the attribute. Each attribute can have one or multiple populators. When several populators are used, they are chained. There are four types of populator:

1) *Numeric values populator*: This populator, `@NumberValue`, serves to assign a numerical value to its attribute. Each generator is using the most generic data type it can (`double` for continuous distributions and `long` for discrete distributions) and the populator is responsible for transforming the value to the attribute data type.

2) *Text value populator*: This populator serves to change the provided value to `String`. It can be used on numeric values or on objects, when appropriate `toString()` method is invoked. When optional parameter `length` is set, the string will be trimmed to the required length.

3) *Array populator and collection populator*: This populator serves to populate the provided array with values generated by a primitive value generator. Its parameters allows to specify the minimal and maximal size of the array (array of random length will be created) or exact length of the array. It is also possible to specify target type, to which the generated value will be casted, which is usefull when array is declared for an interface or an abstract class.

D. Populator chaining

Multiple populators can be specified for each attribute. This can be usefull for example when the value has to be transformed from number to string and then used to populate an array. Unfortunately, Java does not guarantee the order in which annotations will be processed, so we had to add `@PropertyPopulatorOrder` annotation, that will define an explicit order in which populators are used.

VI. CASE STUDY

The tool was tested in a real stress testing and benchmarking of the system for receipt processing. The system is composed

of central server collecting and distributing data for a large number of cash registers. It is used for distributing information about product price and also for mandatory electronic record of sales and receipt confirmation.

A. System Setup

The system consists of the central server, divided to application and database part that should serve multiple different companies, with a separate database for each company.

The goal of the measurement was to investigate how high load the server can handle, how fast it will be able to respond and also how it will behave under high load. The testing was performed on the production server and, with multiple instances of clients, modified to use automatically generated data instead of working with user input. The servers were running on Dell PowerEdge R820 cluster in a virtual environment, each server with 4 cores and 4 GB RAM, the database servers were equipped with Postgres database 9.5, each had 2 cores and 16 GB RAM. The clients were launched in bulks of 100 clients on one computer with Intel Xeon E3-1246, with 4 cores and 16 GB RAM. Clients were randomly divided into groups of 1 to 10 clients to simulate different size of companies with more cash registers — each group of clients shares one database. The environment of clients was observed during the whole duration of test to ensure that the clients will not become the bottleneck of the test. Eight instances of the test was run, with increasing number of clients.

B. Measured results

The experiments were running 24 hours, with a constant setting simulating high business load. The results are summarized in the table I. The table shows average and median times both for the processing of receipt and for processing of update, each time is measured from the start of the operation till its final confirmation. The measurement excluded the time required for data generating. The distribution of response times for updates and receipts was close to the exponential distribution, as can be demonstrated on histograms created for fourth experiment (other experiments showed similar behaviour). Stress tests helped to find several problems in server and client implementation, most notably with server side memory management and with reaction of clients to failed updates.

C. Experience with generator tool

The testing was performed on the application in several stages of the development. As the developers relied on agile methodology, source texts of the application went through several changes, not only regarding the added features, but also in a structure of the domain modal. Most notably, the `Receipt` class changed several times, from simple class with several attributes and two arrays with names and prices of items to an aggregated class containing collections of `Item` instances, customer and seller identification and other complex attributes. As the test data were generated automatically, these changes required only to add further annotations to the

TABLE I
EXPERIMENT SETUP AND MEASURED RESULTS

experiment	Setup		Updates			Receipts			
	clients	avg. time [ms]	med. time [ms]	SD	avg. size [B]	avg. time [ms]	med. time [ms]	SD	avg. size [B]
1	200	12.0	5.7	22.0	178300	5.1	3.8	4.3	35833
2	400	11.3	5.9	19.9	177200	6.7	4.2	9.0	36920
3	600	57.2	12.5	74.4	171700	6.5	4.9	4.3	35750
4	800	77.7	30.6	82.4	166800	8.9	6.3	6.9	36160
5	1000	146.5	159.6	116.8	174600	10.2	6.8	8.8	37210
6	1200	98.9	70.6	91.3	172900	17.2	11.4	16.4	36330
7	1500	147.1	183.7	116.2	180250	33.6	23.7	30.6	36140
8	2000	140.2	140.9	84.3	184650	101.8	92.9	79.3	36780

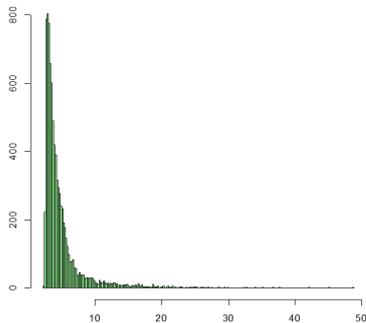


Fig. 3. Histogram of time (in milliseconds) required to perform update, normalized according to size.

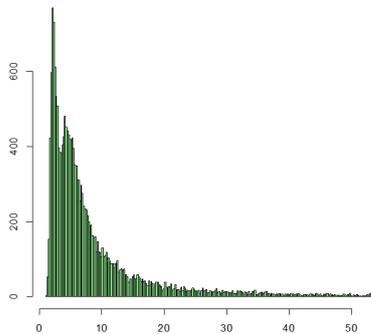


Fig. 4. Histogram of time (in milliseconds) required to accept receipt, normalized according to size.

domain objects, without need to make any changes in the tests themselves.

In order to see how demanding the running of the data generators during testing is, we measured times required to generate instances of data classes. In Table II you can see the times required for generating one instance of `Item` and `Receipt` classes. In last row the time required to generate one instance of the `Receipt` class with 10 `Items` is shown. The first column shows average duration of the first run of the generator, the second column shows average time required when all JVM optimization took effect, after generating 100 of instances.

TABLE II
TIMES REQUIRED FOR DATA GENERATING

Class	first generating [ms]	after optimization [ms]
Item	9.25	1.12
Receipt	12.37	1.82
full Receipt	123.15	10.51

VII. RELATED WORK AND COMPARISON TO OTHER TOOLS

Testing based on random data is well established both in literature and practice. In some cases, the generators are very complex. For example in [4] a random generator is used to test the compiler by randomly generating Pascal programs fulfilling provided grammar and testing many possible paths of compilation. Several other examples of using random generators are summarized in [5]. The random generators are also used to achieve better code coverage in testing in different setups for a long time - [6], [7], [8], often in combination with some technique to limit the number of generated test data. In [6], genetic algorithms are used to search for test data that provides the best code coverage and such approaches are still investigated [9], [?]. In [8], random test data are used in combination with guidance obtained from runtime analysis of the program under test. Similar approaches used not only for unit tests or load tests, but also for the tests of the user interfaces [10]. However, most literature focused on testing deals with methods for creating data in a deterministic way, in order to maximize code coverage of unit tests.

Several tools that allow generating of complex test data exists, but they are mostly intended for use in web applications or to test web services [11], [12], [13]. Typically, they allow to generate datasets in formats like JSON or XML, and use them as a result of web service or input for further processing. The definition of data structure is then separated from the rest of the program and tests, so it cannot dynamically react to changes in definition of data carrying objects. The main difference is in the ability to work with complex structures. The mentioned generators are able to generate test data according to the defined data structures, but cannot analyze the domain

objects that are connected with the transferred data and cannot accommodate to its changes.

The closest tool similar to JOP is PODAM [14], POJO DATA Mocker. This tool allows analysis of POJO objects and fill them with random data. It also supports both primitive data types and work with Java collections, user factories for supplying data that cannot be generated directly and additional execution of objects methods when data generation is finished. The mocker is designed to work in Spring framework environment, so when the application under test should use the PODAM generator, it becomes dependent on large part of the Spring framework. The main difference between PODAM and JOP is limited support of PODAM of working with references between objects - PODAM basically support only generation of additional objects in the object tree and no searching between already generated objects.

VIII. CONCLUSION AND FUTURE WORK

We have presented a tool that should simplify testing, when a large amount of testing data object is required. Although such tools already exist, we believe that our approach helps to make tests simpler, by moving the definition of the data structure from the tests to the classes. This way, source texts of the tests are more independent on the implementation of data carrying classes. The data structure is kept on one clearly defined place. The tool is intended mainly for stress tests, measuring of quality of services and integration tests than for unit tests, as the generators are focused on creating of realistic looking data and not to achieve maximal code coverage. The generating is fully automated and does not require any effort from the tester, however, it requires the creator of data carrying classes to create specifications of the data structure.

Currently, we have the prototype implementation of the described tool, we are now working on creating a distributable version. The implementation is available publicly on GitHub, along with set of basic examples¹. We have several issues that need to be addressed to make the tool more useful.

Our main focus is now to modify the JOP to allow the generating unit test data that would ensure the code coverage of the application under test. The current implementation is focused mainly on the stress testing and thus creating large number of data, but it seems useful to be able to direct the generating algorithm to the critical points of the application. As the code coverage achieved through the different data sets can lead to an enormous number of test cases, we also experiment with approaches to minimize the size of the test set using methods of combinatorial testing [15] and particle swarm optimization [16].

The other thing we would like to focus on is to adapt the description of the data generators to the form of the constraint of each attribute. Such constraints can then have a wider use, for example for validation of user input.

ACKNOWLEDGMENT

The authors would like to thank Michal Dekany, who did a great job implementing the ideas from the paper to the working tool.

This work was supported by Ministry of Education, Youth and Sports of the Czech Republic, Institutional support for long-term strategic development of research organizations.

REFERENCES

- [1] D. W. Dyer. Uncommon math. Accessed: 2018-05-05. [Online]. Available: <http://maths.uncommons.org/>
- [2] Random string utils. Accessed: 2018-05-05. [Online]. Available: <https://commons.apache.org/proper/commons-lang/javadocs/api-2.6/org/apache/commons/lang/RandomStringUtils.html>
- [3] P. Brémaud, *Markov chains : Gibbs fields, Monte Carlo simulation and queues*, ser. Texts in applied mathematics. New York, Berlin, Heidelberg: Springer, 1999. ISBN 0-387-98509-3. [Online]. Available: <http://opac.inria.fr/record=b1094914>
- [4] F. Bazzichi and I. Spadafora, "An automatic generator for compiler testing," *IEEE Trans. Softw. Eng.*, vol. 8, no. 4, pp. 343–353, Jul. 1982. doi: 10.1109/TSE.1982.235428. [Online]. Available: <http://dx.doi.org/10.1109/TSE.1982.235428>
- [5] B. Wichmann. Some Remarks About Random Testing. Accessed: 2018-05-05. [Online]. Available: http://www.npl.co.uk/upload/pdf/random_testing.pdf
- [6] C. C. Michael, G. E. McGraw, M. A. Schatz, and C. C. Walton, "Genetic algorithms for dynamic test data generation," in *Automated Software Engineering, 1997. Proceedings., 12th IEEE International Conference*, Nov 1997. doi: 10.1109/ASE.1997.632858 pp. 307–308.
- [7] S. Poulding and J. A. Clark, "Efficient software verification: Statistical testing using automated search," *IEEE Transactions on Software Engineering*, vol. 36, no. 6, pp. 763–777, Nov 2010. doi: 10.1109/TSE.2010.24
- [8] L. Ma, C. Artho, C. Zhang, H. Sato, J. Gmeiner, and R. Ramler, "Grt: An automated test generator using orchestrated program analysis," in *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, Nov 2015. doi: 10.1109/ASE.2015.102 pp. 842–847.
- [9] C. Koleejan, B. Xue, and M. Zhang, "Code coverage optimisation in genetic algorithms and particle swarm optimisation for automatic software test data generation," in *2015 IEEE Congress on Evolutionary Computation (CEC)*, May 2015. doi: 10.1109/CEC.2015.7257026. ISSN 1089-778X pp. 1204–1211.
- [10] K. Salvesen, J. P. Galeotti, F. Gross, G. Fraser, and A. Zeller, "Using dynamic symbolic execution to generate inputs in search-based gui testing," in *2015 IEEE/ACM 8th International Workshop on Search-Based Software Testing*, May 2015. doi: 10.1109/SBST.2015.15 pp. 32–35.
- [11] Mockaroo. Accessed: 2018-05-05. [Online]. Available: <https://www.mockaroo.com/>
- [12] Dtm test xml generator. Accessed: 2018-05-05. [Online]. Available: <http://www.sqledit.com/xmlgenerator/>
- [13] Redgate. Accessed: 2018-05-05. [Online]. Available: <http://www.red-gate.com/products/sql-development/sql-data-generator/>
- [14] Podam - pojo data mocker. Accessed: 2018-05-05. [Online]. Available: <https://github.com/mtdone/podam>
- [15] M. Bures and B. S. Ahmed, "On the effectiveness of combinatorial interaction testing: A case study," in *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, July 2017. doi: 10.1109/QRS-C.2017.20 pp. 69–76.
- [16] B. S. Ahmed, L. M. Gambardella, W. Afzal, and K. Z. Zamli, "Handling constraints in combinatorial interaction testing in the presence of multi objective particle swarm and multithreading," *Information and Software Technology*, vol. 86, pp. 20 – 36, 2017. doi: <https://doi.org/10.1016/j.infsof.2017.02.004>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0950584917301349>

¹<https://github.com/mrfranta/jop/>

Assertional Reasoning for Concurrent and Communicating BPEL-like Programs

Longfei Zhu
Key Laboratory of
Ministry of Public Security,
Zhejiang Police College,
Zhejiang, China
Email: zhulongfei@zjcxxy.cn

Qiwen Xu [§]
Faculty of Science and Technology,
University of Macau,
Macao SAR, China
Email: qwxu@umac.mo

Huibiao Zhu
Shanghai Key Laboratory of
Trustworthy Computing,
East China Normal University,
Shanghai, China
Email: hbzhu@sei.ecnu.edu.cn

Abstract—This paper studies verification of programs similar to BPEL4WS (BPEL), the latter being a de facto standard for the web services composition and orchestration. Traditionally, in verification of concurrent and distributed programs, the model was either based on shared variables or message passing and was studied separately. BPEL-like programs have features that are present in both models: several flows within one service can be executed in parallel and they can access shared variables, whereas several services communicate by message passing. Therefore, it is natural that for verification of BPEL-like programs, the verification methods developed for shared variables and message passing be integrated. In this paper, we combine the proof rules for shared variable programs from Owicki and Gries, the proof rules for CSP like programs from Apt, Francez and de Roever, together with proof rules for compensation and fault handling, to cover all major features of BPEL. An operational semantics is presented and the proof rules can be justified over that. Examples are provided to show the feasibility of verification framework.

Index Terms—BPEL, Hoare logic, shared variables, message passing.

I. INTRODUCTION

WEB services and other web-based applications have been becoming more and more important in practice. Various web-based business process languages have been introduced, such as XLANG [1], WSFL [2], BPEL4WS (BPEL) [3] and StAc [4], which are designed for the description of services composed of a set of processes across the Internet. Their goal is to achieve the universal interoperability between applications by using web standards, as well as to specify the technical infrastructure for carrying out business transactions. BPEL4WS (BPEL) is the OASIS standard for web services composition and orchestration. To support long-running transactions, it provides the ability to define fault and compensation handling. In addition, BPEL allows several flows executing in parallel in a service, and several services running concurrently. Due to the interesting features of BPEL programs mentioned above, the verification of BPEL programs is challenging.

Much research has been done on verification of concurrent and distributed programs. Typically, the model is either based on shared-variables or message passing. Owicki and Gries [5], and Apt, Francez and de Roever [6], respectively extended

Hoare logic to concurrent programs with shared variables and distributed programs with message passing. BPEL-like programs have features that are present in both models: several parallel flows within one service can access shared variables, whereas several concurrent services communicate by message passing. Therefore, it is natural that for verification of BPEL like programs, the verification methods developed for shared variables and message passing be integrated. In this paper, we combine the proof rules for shared variable programs from Owicki and Gries, the proof rules for CSP like programs from Apt, Francez and de Roever, together with proof rules for compensation and fault handling, to cover all major features of BPEL.

The remainder of this paper is organized as follows. Section 2 introduces a language based on BPEL together with an operational semantics. In section 3, we provide the verification rules, including the rules for dealing with compensation, fault handling, parallel flows through shared variables, and multiple services through message passing. A few simple examples are given to illustrate the rules. Section 4 concludes the paper with a discussion.

II. AN OPERATIONAL MODEL

In this section, we present the operational semantics of a BPEL-like language, based on the work in [7] and [8].

A. The Syntax of BPEL

Our language contains the following categories of syntactic elements:

$$\begin{aligned} BA & ::= \text{skip} \mid x := e \mid \text{rec } a \ x \mid \text{rep } a \ e \mid \text{throw} \\ A & ::= BA \mid g \circ A \mid A; A \mid A \triangleleft b \triangleright A \mid b * A \\ & \quad \mid A \parallel A \mid A \sqcap A \mid \text{undo } n \mid \{A? A, A\}_n \\ W & ::= (A, \dots, A) \end{aligned}$$

where:

- The category BA stands for the basic activity. Activity $x := e$ assigns the value of e to variable x . Activity skip behaves the same as $x := x$. A variable may be shared among parallel flows within one service.

In order to implement the communications between concurrent services, two statements are introduced, i.e.,

[§] Qiwen Xu is corresponding author.

$\text{rec } a \ x$ and $\text{rep } a \ e$. Activity $\text{rec } a \ x$ represents the receiving of a value through channel a and storing in x . To avoid complications, we assume variable x is not shared by parallel flows. If the information is needed by another flow, it has to be copied to another variable first. Sending a message is represented by $\text{rep } a \ e$. Activity throw indicates that the program encounters a fault.

- The category A stands for the activities within one service. Several constructs are similar to those in traditional programming languages. $A; B$ stands for sequential composition. $A \triangleleft b \triangleright B$ is the conditional construct and $b * A$ is the iteration construct. $A \sqcap B$ stands for the nondeterministic choice. $g \circ A$ awaits the Boolean condition g to be set true. $\{A?C, F\}_n$ stands for the scope based compensation statement, where n stands for the scope name, A , C and F for the primary activity, compensation program and fault handler correspondingly. If A terminates successfully, program C is installed in the compensation list for later compensating. On the other hand, if A encounters a fault during its execution, the fault handler F will be activated. Further, the compensation part C does not contain scope activity. Statement “undo n ” activates the execution of the programs with scope name n .

A service may contain one or several flows running in parallel. We use the notation $A \parallel B$ to stand for two such flows.

- The category W stands for the coordination of several concurrent services. Such a set of services is denoted by (A_1, \dots, A_n) , and their communication is modelled by message passing.

B. An Operational Model

For the operational semantics of BPEL, its transitions are of the two types.

$$C \longrightarrow C' \quad \text{or} \quad C \xrightarrow{a.m} C'$$

where C and C' are the configurations describing the states of an execution mechanism before and after a step respectively. The first type is used to denote non-communication transitions. The second type is used to represent communication between concurrent services where a is the channel and m is the message that is passed.

A configuration is expressed as $\langle P, \sigma, Cp \rangle$, where

- (1) The first component P is a program that remains to be executed.
- (2) The second element σ is the state for all the variables.
- (3) The third element Cp stands for a compensation set; i.e., containing the scope names whose compensation parts need to be executed. Cp can contain several copies of the same element. Therefore, it is in fact a bag. For a scope n , the compensation program is denoted by $C(n)$. When statement $\text{undo } n$ is executed, $C(n)$ will be invoked.

For the program P in configuration $\langle P, \sigma, Cp \rangle$, it can either be a normal program or one of the following special forms:

ε : A program has terminated successfully. We use ε to represent the empty program.

\boxtimes : A program has encountered a fault and stops at the faulty state, represented by a special symbol \boxtimes .

C. Transition Rules

Transition rules are presented below.

(1) Basic Commands

Firstly we list the operational semantics for basic commands. The execution of $x := e$ assigns the value of expression e to variable x , and leaves other variables unchanged.

$$\langle x := e, \sigma, Cp \rangle \longrightarrow \langle \varepsilon, \sigma[x \mapsto e(\sigma)], Cp \rangle$$

For communication commands, statement $\text{rec } a \ x$ receives message m through channel a . The received message will be stored in variable x .

$$\langle \text{rec } a \ x, \sigma, Cp \rangle \xrightarrow{a.m} \langle \varepsilon, \sigma[x \mapsto m], Cp \rangle$$

$\text{rep } a \ e$ stands for the sending of e on channel a , and the message is $e(\sigma)$ when sent in state σ .

$$\langle \text{rep } a \ e, \sigma, Cp \rangle \xrightarrow{a.e(\sigma)} \langle \varepsilon, \sigma, Cp \rangle$$

throw encounters a fault after activation, while leaving all variables and the compensation set unchanged.

$$\langle \text{throw}, \sigma, Cp \rangle \longrightarrow \langle \boxtimes, \sigma, Cp \rangle$$

$\text{undo } n$ invokes the compensation program corresponding to scope name n .

$$\langle \text{undo } n, \sigma, Cp \rangle \longrightarrow \langle C(n), \sigma, Cp \setminus n \rangle, \quad \text{where } n \in Cp$$

Here function $C(n)$ represents the program whose name is n (i.e, the scope name). $Cp \setminus n$ represents that scope name n is removed once from Cp .

(2) Sequential Constructs

For sequential composition $P; Q$, if P does not encounter a fault, the transition rules are the same as usual. Below in this section, $\xrightarrow{\beta}$ denotes either a communication or non-communication transition.

$$\frac{\langle P, \sigma, Cp \rangle \xrightarrow{\beta} \langle P', \sigma', Cp' \rangle \text{ and } P' \neq \varepsilon, \boxtimes}{\langle P; Q, \sigma, Cp \rangle \xrightarrow{\beta} \langle P'; Q, \sigma', Cp' \rangle}$$

$$\frac{\langle P, \sigma, Cp \rangle \xrightarrow{\beta} \langle \varepsilon, \sigma', Cp' \rangle}{\langle P; Q, \sigma, Cp \rangle \xrightarrow{\beta} \langle Q, \sigma', Cp' \rangle}$$

If P encounters a fault during its execution, $P; Q$ also encounters a fault during its execution.

$$\frac{\langle P, \sigma, Cp \rangle \xrightarrow{\beta} \langle \boxtimes, \sigma', Cp' \rangle}{\langle P; Q, \sigma, Cp \rangle \xrightarrow{\beta} \langle \boxtimes, \sigma', Cp' \rangle}$$

The usual await statement $g \circ P$ waits for the Boolean guard g to be set true.

$$\langle g \circ P, \sigma, Cp \rangle \longrightarrow \langle P, \sigma, Cp \rangle, \text{ if } g(\sigma)$$

$P \sqcap Q$ either behaves like P or like Q . The choice between

them is nondeterministic.

$$\begin{aligned} \langle P \sqcap Q, \sigma, Cp \rangle &\longrightarrow \langle P, \sigma, Cp \rangle \\ \langle P \sqcap Q, \sigma, Cp \rangle &\longrightarrow \langle Q, \sigma, Cp \rangle \end{aligned}$$

The conditional $P \triangleleft b \triangleright Q$ starts process P if the value of b is true. Otherwise it executes Q instead.

$$\begin{aligned} \langle P \triangleleft b \triangleright Q, \sigma, Cp \rangle &\longrightarrow \langle P, \sigma, Cp \rangle, \text{ if } b(\sigma) \\ \langle P \triangleleft b \triangleright Q, \sigma, Cp \rangle &\longrightarrow \langle Q, \sigma, Cp \rangle, \text{ if } \neg b(\sigma) \end{aligned}$$

The transition rules for iteration are similar to conditional.

$$\begin{aligned} \langle b * P, \sigma, Cp \rangle &\longrightarrow \langle P; b * P, \sigma, Cp \rangle, \text{ if } b(\sigma) \\ \langle b * P, \sigma, Cp \rangle &\longrightarrow \langle \varepsilon, \sigma, Cp \rangle, \text{ if } \neg b(\sigma) \end{aligned}$$

(3) Parallel Flows

Now we consider the transition rules for parallel composition. First we define a function $\mathbf{par}(P, Q)$, which can be used in defining the transition rules for parallel composition. Let

$$\mathbf{par}(P, Q) =_{df} \begin{cases} \varepsilon & \text{if } P = \varepsilon \wedge Q = \varepsilon \\ \boxtimes & \text{if } P = \boxtimes \wedge Q = \boxtimes \\ \vee P = \boxtimes \wedge Q = \varepsilon \\ \vee P = \varepsilon \wedge Q = \boxtimes \\ P \parallel Q & \text{otherwise} \end{cases}$$

It indicates the program status for two parallel flows after executing a transition. If both components are in the empty states, the whole service is also in the empty state. If both are in the faulty states, or one is in the faulty state and another one is in the empty state, then the whole service is also in faulty state. If one flow performs a transition, the whole service can also perform the transition.

$$\begin{aligned} &\frac{\langle P, \sigma, Cp \rangle \xrightarrow{\beta} \langle P', \sigma', Cp' \rangle}{\langle P \parallel Q, \sigma, Cp \rangle \xrightarrow{\beta} \langle \mathbf{par}(P', Q), \sigma', Cp' \rangle} \\ &\frac{\langle Q, \sigma, Cp \rangle \xrightarrow{\beta} \langle Q', \sigma', Cp' \rangle}{\langle P \parallel Q, \sigma, Cp \rangle \xrightarrow{\beta} \langle \mathbf{par}(P, Q'), \sigma', Cp' \rangle} \end{aligned}$$

(4) Scope

For scope $\{A?C, F\}_n$, if the primary activity A performs a transition which does not lead to the faulty state, the whole scope can also perform the successful transition of the same type.

$$\frac{\langle A, \sigma, Cp \rangle \xrightarrow{\beta} \langle A', \sigma', Cp' \rangle \text{ and } A' \neq \boxtimes}{\langle \{A?C, F\}_n, \sigma, Cp \rangle \xrightarrow{\beta} \langle \{A'?C, F\}_n, \sigma', Cp' \rangle}$$

When the primary activity has been terminated, the compensation program is added into the compensation set. This is represented by the following rule.

$$\langle \{\varepsilon?C, F\}_n, \sigma, Cp \rangle \longrightarrow \langle \varepsilon, \sigma, Cp \cup \{n \rightarrow C\} \rangle$$

On the other hand, if the primary activity performs a transition leading to the faulty state, the fault handler in the scope will

be activated.

$$\frac{\langle A, \sigma, Cp \rangle \xrightarrow{\beta} \langle \boxtimes, \sigma', Cp' \rangle}{\langle \{A?C, F\}_n, \sigma, Cp \rangle \xrightarrow{\beta} \langle F, \sigma', Cp' \rangle}$$

(5) Communicating Services

A collection of concurrent services is represented as $W = (P_1, P_2, \dots, P_n)$, and we use σ_i and Cp_i to denote the state and compensation set of service P_i respectively.

If one service does non-communication transitions, the whole system can also do a transition of the same type.

$$\frac{\langle P_i, \sigma_i, Cp_i \rangle \longrightarrow \langle P'_i, \sigma'_i, Cp'_i \rangle}{\langle W, \sigma, Cp \rangle \longrightarrow \langle W', \sigma', Cp' \rangle}$$

where $W' = (P_1, P_2, \dots, P'_i, \dots, P_n)$, $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_i, \dots, \sigma_n)$, $\sigma' = (\sigma_1, \sigma_2, \dots, \sigma'_i, \dots, \sigma_n)$ and $Cp' = (Cp_1, Cp_2, \dots, Cp'_i, \dots, Cp_n)$.

If two services involve the communication via the same channel, the whole system also does the communication via the channel.

$$\frac{\langle P_i, \sigma_i, Cp_i \rangle \xrightarrow{a.m} \langle P'_i, \sigma'_i, Cp'_i \rangle, \langle P_j, \sigma_j, Cp_j \rangle \xrightarrow{a.m} \langle P'_j, \sigma'_j, Cp'_j \rangle}{\langle W, \sigma, Cp \rangle \longrightarrow \langle W', \sigma', Cp' \rangle}$$

where $W' = (P_1, \dots, P'_i, \dots, P'_j, \dots, P_n)$, $\sigma' = (\sigma_1, \dots, \sigma'_i, \dots, \sigma'_j, \dots, \sigma_n)$, $Cp' = (Cp_1, \dots, Cp'_i, \dots, Cp'_j, \dots, Cp_n)$.

III. VERIFICATION RULES

In this section, we study the verification rules for the BPEL-like programs.

A. Correctness Formula

The verification rules are in the form of a Hoare triple:

$$\{p\} S \{q\}$$

here S stands for the program, p and q stand for the precondition and the postcondition respectively. If the program S is started in a state that satisfies p , after the execution, postcondition q should be satisfied.

To deal with the two typical features of BPEL, i.e., fault handling and compensation, we introduce two variables ok and $comp$.

B. General Rules

Boolean variable ok is used to identify whether a program is in the faulty state or not. For a configuration $\langle P, \sigma, Cp \rangle$, ok is true if and only if $P \neq \boxtimes$. Since the initial configuration is never faulty, we have the following general rule

OK-rule

$$\frac{\{p \wedge ok\} S \{q\}}{\{p\} S \{q\}}$$

ok may be false in the postcondition, indicating that the current system has encountered faults in the execution.

The other general rule is the usual consequence rule:
Consequence-rule

$$\frac{p \Rightarrow p_1, \{p_1\} S \{q_1\}, q_1 \Rightarrow q}{\{p\} S \{q\}}$$

C. Rules for Basic Commands

(1) Assignment:

The rule for assignment is the same as in the traditional Hoare logic and *ok* is true in the postcondition.

$$\{p[e/x]\} x := e \{p \wedge ok\}$$

(2) throw:

For `throw`, it immediately enters into the faulty state while leaving the states unchanged.

$$\frac{r \text{ does not contain variable } ok}{\{r\} \text{throw} \{-ok \wedge r\}}$$

To verify communicating processes, Apt, Francez and de Roever [6] suggested the verification be divided into two phases. The first phase is the “local verification” for each process, and the second phase is the “cooperation test” where the local verification of the processes are checked to be matching.

(3) Replying:

Obviously, sending a message does not change the state

$$\{p\} \text{rep } a \ y \{p\}$$

ok actually holds in the postcondition, but we can deduce this fact by applying the OK-rule.

(4) Receiving:

$$\frac{q \Rightarrow ok}{\{p\} \text{rec } a \ x \{q\}}$$

This rule at first would look odd, as the postcondition can be anything (in our context, as long as *ok* is true). Whether the postcondition is really valid is checked in the cooperation test.

The rule for conditional choice is the same as the traditional one.

(5) Conditional choice:

$$\frac{\{p \wedge b\} S_1 \{q\}, \{p \wedge \neg b\} S_2 \{q\}}{\{p\} S_1 \triangleleft b \triangleright S_2 \{q\}}$$

(6) Sequential Composition

For sequential composition, there are two rules. The first rule stands for the case that the first program successfully terminates. The second rule indicates that the first program encounters fault during its execution.

Rule 1:

$$\frac{r \Rightarrow ok, \{p\} A \{r\}, \{r\} B \{q\}}{\{p\} A ; B \{q\}}$$

Rule 2:

$$\frac{r \Rightarrow \neg ok, \{p\} A \{r\}}{\{p\} A ; B \{r\}}$$

(7) Iteration

For simplicity, we only present the rules for partial correctness.

Rule 1:

$$\frac{\{p \wedge b\} S \{p\}}{\{p\} \text{while } b \text{ do } S \{p \wedge \neg b\}}$$

Rule 2:

$$\frac{q \Rightarrow \neg ok, \{p \wedge b\} S \{q\}}{\{p\} \text{while } b \text{ do } S \{q\}}$$

D. Scope and Compensation

A compensation may be installed several times, so we introduce a function *comp* to record that. More specifically, for a scope *n*, we use *comp.n* to stand for the number that the compensation program has been installed. For the compensated program named *n*, we use function *C(n)* to represent it.

For scope, the verification rules are divided into two cases.

(1) Scope

The first rule deals with the case that the primarily activity *A* can successfully terminate. The compensation program *C* is installed.

Rule 1:

$$\frac{\{p\} A \{q[comp.n + 1/comp.n]\}, q \Rightarrow ok}{\{p\} \{A?C, F\}_n \{q\}}$$

The second rule handles the case that *A* encounters the fault. The fault handler will be triggered.

Rule 2:

$$\frac{\{p\} A \{r \wedge \neg ok\}, \{r\} F \{q\}}{\{p\} \{A?C, F\}_n \{q\}}$$

(2) Compensation

For `undo n`, the compensation program *C(n)* will be executed. In addition, it has the effect of reducing *comp.n* by 1. Therefore, in the precondition of *C(n)*, the number of the recorded program named *n* should be one less.

$$\frac{\{p[comp.n + 1/comp.n]\} C(n) \{q\}}{\{p\} \text{undo } n \{q\}}$$

Example 1 Consider the program below.

```
{x := x + 1?x := x - 1, skip}_n ;
{x := x + 2?x := x - 2, skip}_m ;
undo m;
undo n
```

By applying the verification rules, we can obtain the following proof outline:

```
{ok ∧ x = 0 ∧ comp.n = 0 ∧ comp.m = 0}
{x := x + 1?x := x - 1, skip}_n ;
{ok ∧ x = 1 ∧ comp.n = 1 ∧ comp.m = 0}
{x := x + 2?x := x - 2, skip}_m ;
```

$$\{ok \wedge x = 3 \wedge comp.n = 1 \wedge comp.m = 1\}$$

undo m ;

$$\{ok \wedge x = 1 \wedge comp.n = 1 \wedge comp.m = 0\}$$

undo n

$$\{ok \wedge x = 0 \wedge comp.n = 0 \wedge comp.m = 0\}$$

in which the verification of undo m is supported by

$$\{ok \wedge x = 3 \wedge comp.n = 1 \wedge comp.m + 1 = 1\}$$

$x := x - 2$

$$\{ok \wedge x = 1 \wedge comp.n = 1 \wedge comp.m = 0\}$$

and of undo n by

$$\{ok \wedge x = 1 \wedge comp.n + 1 = 1 \wedge comp.m = 0\}$$

$x := x - 1$

$$\{ok \wedge x = 0 \wedge comp.n = 0 \wedge comp.m = 0\}$$

In this example, the compensation programs completely undo the effect of the forward activities, so it should be expected that final postcondition is exactly the same as the initial precondition.

E. Parallel Flows

In one service, several flows may be executed in parallel and information is exchanged via shared variables. In the classic verification method due to Owicki and Gries [5], the central concept is the interference freedom. Intuitively, it means that assertions in the local proofs of one process should not be invalidated by the execution of a parallel process. Suppose $\{p\}S\{q\}$ is a Hoare triple in the local verification for the statement S , statement T from another process is said to be interference free to $\{p\}S\{q\}$ if the following two conditions are satisfied:

- (1) $\{\exists ok.p \wedge pre(T)\}T\{\exists ok.p\}$
- (2) $\{\exists ok.q \wedge pre(T)\}T\{\exists ok.q\}$

where $pre(T)$ is the precondition of T . Note the interference freedom is concerned with the shared program variables, and hence ok is removed from the assertions by the quantification. Adopting the parallel rule to our setting, the postcondition is modified to take into account the faulty states.

$$\frac{\{p_i\}S_i\{q_i\} \text{ are interference-free}}{\{p_1 \wedge p_2\}S_1 \parallel S_2\{Merge(q_1, q_2)\}}$$

where $Merge(q_1, q_2) =_{df} \exists ok_1, ok_2 \bullet q_1[ok_1/ok] \wedge q_2[ok_2/ok] \wedge ok = ok_1 \wedge ok_2$. This combines the two postconditions, for the information about local variables and compensation, and the parallel flow is in the faulty state if at least one component is in the faulty state.

Example 2 Let $S_1 =_{df} x := x + 1 ; \text{throw}$, $S_2 =_{df} x := x + 2$.

For S_1 , we have the following local proof outline

$$\{x = 0\}$$

$$\{ok \wedge (x = 0 \vee x = 2)\}$$

$x := x + 1$

$$\{ok \wedge (x = 1 \vee x = 3)\}$$

throw

$$\{\neg ok \wedge (x = 1 \vee x = 3)\}$$

For S_2 , the proof outline is

$$\{x = 0\}$$

$$\{ok \wedge (x = 0 \vee x = 1)\}$$

$x := x + 2$

$$\{ok \wedge (x = 2 \vee x = 3)\}$$

For interference freedom test, we need to check assertions $(x = 0 \vee x = 2)$ and $(x = 1 \vee x = 3)$ in S_1 are not invalidated by $x := x + 2$ in S_2 , whereas $(x = 0 \vee x = 1)$ and $(x = 2 \vee x = 3)$ in S_2 are not invalidated by $x := x + 1$ in S_1 . Formally, this is shown by the following

$$\begin{aligned} &\{(x = 0 \vee x = 2) \wedge (x = 0 \vee x = 1)\} x := x + 2 \{x = 0 \vee x = 2\} \\ &\{(x = 1 \vee x = 3) \wedge (x = 0 \vee x = 1)\} x := x + 2 \{x = 1 \vee x = 3\} \\ &\{(x = 0 \vee x = 1) \wedge (x = 0 \vee x = 2)\} x := x + 1 \{x = 0 \vee x = 1\} \\ &\{(x = 2 \vee x = 3) \wedge (x = 0 \vee x = 2)\} x := x + 1 \{x = 2 \vee x = 3\}, \end{aligned}$$

which are all trivial. By the rule for parallel flows, we have

$$\{x = 0\}S_1 \parallel S_2\{\neg ok \wedge x = 3\}$$

F. Communicating Services

Different services do not share variables and communicate by passing messages. The central concept in the method developed by Apt, Francez and de Roever [6] is the cooperation test. It checks that the postcondition of an input command is indeed ensured by the sending command. The Hoare triples of two matching communication pairs

$$\begin{aligned} &\{p_1\} \text{rec } a \ x \ {q_1} \\ &\{p_2\} \text{rep } a \ e \ {q_2} \end{aligned}$$

cooperate, if the following is true

$$\{\exists ok. p_1 \wedge p_2\} x := e \{\exists ok. q_1 \wedge q_2\}$$

For a set of services, the proof outlines cooperate if the Hoare triples of every two matching communication pairs does. Note the assertions in the verification of each service may contain ok and $comp$, and we rename them as ok_i and $comp_i$ to avoid conflicts among different services, and arrive at the following rule for services

$$\frac{\text{proof of } \{p_i\}P_i\{q_i\} \text{ cooperate, } i = 1, 2, \dots, n}{\{p_1 \wedge p_2 \wedge \dots \wedge p_n\}(P_1, P_2, \dots, P_n)\{q'_1 \wedge q'_2 \wedge \dots \wedge q'_n\}}$$

where $q'_i = q_i[ok_i/ok, comp_i/comp]$.

Example 3 Let $P_1 =_{df} x_1 := 0; \text{rep } a \ (x_1 + 1)$, $P_2 =_{df} \text{rec } a \ x_2; \text{rep } b \ (x_2 + 2)$, $P_3 =_{df} \text{rec } b \ x_3$.

For P_1 , we have the following proof outline

$$\begin{array}{l} \{\text{true}\} \\ \{ok\} \\ x_1 := 0 \\ \{ok \wedge x_1 = 0\} \\ \text{rep } a (x_1 + 1) \\ \{ok \wedge x_1 = 0\} \end{array}$$

For P_2 ,

$$\begin{array}{l} \{\text{true}\} \\ \{ok\} \\ \text{rec } a x_2 \\ \{ok \wedge x_2 = 1\} \\ \text{rep } b (x_2 + 2) \\ \{ok \wedge x_2 = 1\} \end{array}$$

For P_3 ,

$$\begin{array}{l} \{\text{true}\} \\ \{ok\} \\ \text{rec } b x_3 \\ \{ok \wedge x_3 = 3\} \end{array}$$

There are two matching communication pairs. For cooperation test, we need to check

$$\begin{array}{l} \{x_1 = 0\} \quad x_2 := x_1 + 1 \quad \{x_2 = 1\} \\ \{x_2 = 1\} \quad x_3 := x_2 + 2 \quad \{x_3 = 3\} \end{array}$$

which are all trivial. It follows that

$$\begin{array}{l} \{\text{true}\} \\ (P_1, P_2, P_3) \\ \{ok_1 \wedge ok_2 \wedge ok_3 \wedge x_1 = 0 \wedge x_2 = 1 \wedge x_3 = 3\} \end{array}$$

IV. CONCLUSION

There has been some work on applying formal methods to web services. An operational semantics of StAC (Structured Activity Compensation) [9], another business process modeling language where compensation acts as one of its main features, has also been studied in [4]. StAC and the B method has been combined in [10] to describe business transactions. Bruni *et al.* [11] have studied the transaction calculi for Sagas. The long-running transactions were discussed and a process calculi was proposed in [12] in the context of a Java API, namely the Java Transactional Web Services. Laneve and Zavattaro [13] explored the application of π -calculus in the formalization of the semantics of the transactional construct of BPEL. They also studied a standard pattern of Web Services composition using π -calculus. For verifying the properties of long-running transactions, Lanotte *et al.* [14] have explored their approach in a timed framework, where a Communicating Hierarchical Timed Automata was developed. Model checking techniques have been applied in the verification of properties of long-running transactions.

In comparison, there has been little work on deductive reasoning of BPEL-like programs. As far as we know, Luo *et al.* [15] were the first to study a Hoare logic for BPEL-like programs. The work has not covered concurrent behaviours. Parallelism in one service has been considered in [8], and

the rely/guarantee [16] approach to verifying shared variable programs is adopted. The same approach (instead of rely/guarantee, usually named as assumption/commitment) for message passing, although also available, e.g., see [17] for a survey, is more difficult to use. Therefore, in this paper, we decide to adopt the earlier cooperation test approach from Apt, Francez and de Roever. To be consistent in the style, the method of interference freedom test from Owicki and Gries is adopted to deal with shared variables.

In this paper, we focus on the deductive reasoning of BPEL-like programs in one unified framework, especially the verification of concurrent communicating BPEL programs. Verification methods developed for shared variables and message passing are integrated. To deal with the compensation and fault handling of web services and facilitate the verification, we introduce *ok* and present the corresponding rules. There are a few minor technical improvements over [8] in the way *ok* is used. Examples are provided to show the feasibility of verification framework.

ACKNOWLEDGMENT

This work is supported in part by National Key R&D Program of China (No. 2017YFC0803700), Macao Science and Technology Development Fund under the EAE project (No. 072/2009/A3), Ministry of Public Security of China (No. 2017GABJC16), Natural Science Foundation of China (No. 61602177 and No. 61402176). The authors would like to thank J.W. Sanders, C. Ma and X. Liu for discussions.

REFERENCES

- [1] S. Thatte, *XLANG: Web Service for Business Process Design*. Microsoft, 2001, http://www.gotdotnet.com/team/xml_wsspecs/xlang-c/default.html.
- [2] F. Leymann, *Web Services Flow Language (WSFL 1.0)*. IBM, 2001, <http://www-3.ibm.com/software/solutions/webservices/pdf/WSDL.pdf>.
- [3] F. Curbera, Y. Golland, J. Klein, F. Leymann, D. Roller, M. Satish Thatte, and S. Weerawarana, *Business Process Execution Language for Web Service*, 2003, <http://www.siebel.com/bpel>.
- [4] M. J. Butler and C. Ferreira, "An operational semantics for StAC, a language for modelling long-running business transactions," in *Proc. COORDINATION 2004: 6th International Conference on Coordination Models and Languages, Pisa, Italy, February 24–27, 2004*, ser. Lecture Notes in Computer Science, vol. 2949. Springer-Verlag, 2004, pp. 87–104.
- [5] S. S. Owicki and D. Gries, "An axiomatic proof technique for parallel programs I," *Acta Informatica*, vol. 6, pp. 319–340, 1976. doi: 10.1007/BF00268134. [Online]. Available: <https://link.springer.com/article/10.1007/BF00268134>
- [6] K. R. Apt, N. Francez, and W. P. D. Roever, "A proof system for communicating sequential processes," vol. 2, no. 3, pp. 359–385. doi: 10.1145/357103.357110. [Online]. Available: <http://dl.acm.org/citation.cfm?id=357110>
- [7] Z. Qiu, S. Wang, G. Pu, and X. Zhao, "Semantics of BPEL4WS-Like fault and compensation handling," in *Proc. FM 2005: International Symposium of Formal Methods Europe, Newcastle, UK, July 18–22, 2005*, ser. Lecture Notes in Computer Science, vol. 3582. Springer-Verlag, 2005, pp. 350–365.
- [8] H. Zhu, Q. Xu, C. Ma, S. Qin, and Z. Qiu, "The rely/guarantee approach to verifying concurrent bpel programs," in *Software Engineering and Formal Methods - 10th International Conference, SEFM 2012, Thessaloniki, Greece, October 1–5, 2012. Proceedings*, ser. Lecture Notes in Computer Science, vol. 7504. Springer, 2012. doi: 10.1007/978-3-642-33826-7_12 pp. 172–187. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-33826-7_12

- [9] M. J. Butler and C. Ferreira, "A process compensation language," in *Proc. IFM 2000: 2nd International Conference on Integrated Formal Methods, Dagstuhl Castle, Germany, November 1–3, 2000*, ser. Lecture Notes in Computer Science, vol. 1945. Springer-Verlag, 2000, pp. 61–76.
- [10] M. J. Butler, C. Ferreira, and M. Y. Ng, "Precise modelling of compensating business transactions and its application to BPEL," *Journal of Universal Computer Science*, vol. 11, no. 5, pp. 712–743, 2005.
- [11] R. Bruni, H. C. Melgratti, and U. Montanari, "Theoretical foundations for compensations in flow composition languages," in *Proceedings of the 32Nd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, ser. POPL '05. ACM, doi: 10.1145/1040305.1040323. ISBN 978-1-58113-830-6 pp. 209–220. [Online]. Available: <http://doi.acm.org/10.1145/1040305.1040323>
- [12] R. Bruni, G. L. Ferrari, H. C. Melgratti, U. Montanari, D. Strollo, and E. Tuosto, "From theory to practice in transactional composition of web services," in *Formal Techniques for Computer Systems and Business Processes*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, doi: 10.1007/11549970_20 pp. 272–286. [Online]. Available: https://link.springer.com/chapter/10.1007/11549970_20
- [13] C. Laneve and G. Zavattaro, "Web-pi at work," in *Proc. TGC 2005: International Symposium on Trustworthy Global Computing, Edinburgh, UK, April 7–9, 2005*, ser. Lecture Notes in Computer Science, vol. 3705. Springer-Verlag, 2005, pp. 182–194.
- [14] R. Lanotte, A. Maggiolo-Schettini, P. Milazzo, and A. Troina, "Design and verification of long-running transactions in a timed framework," *Science Computer Programming*, vol. 73, no. 2-3, pp. 76–94, 2008. doi: 10.1016/j.scico.2008.07.001. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167642308000774>
- [15] C. Luo, S. Qin, and Z. Qiu, "Verifying bpel-like programs with hoare logic," in *Proc. TASE 2008: 2nd IEEE International Symposium on Theoretical Aspects of Software Engineering*. Nanjing, China: IEEE Computer Society, June 2008, pp. 151–158.
- [16] C. B. Jones, "Tentative steps toward a development method for interfering programs," *ACM Trans. Program. Lang. Syst.*, vol. 5, no. 4, pp. 596–619, 1983. doi: 10.1145/69575.69577. [Online]. Available: <http://dl.acm.org/citation.cfm?id=69577>
- [17] W.-P. de Roever, F. de Boer, U. Hannemann, J. Hooman, Y. Lakhnech, and M. P. J. Zwiers, *Concurrency Verification: Introduction to Compositional and Noncompositional Methods*. Cambridge Tracts in Theoretical Computer Science 54, Cambridge University Press, 2001.

Author Index

- A**dams, Tim 13
Aleithe, Michael 145
- B**akker, René 87
Banach, Richard 73
Barrett, John 73
Beffa, Corentin 53
Begosso, Luiz Carlos 225
Begosso, Luiz Ricardo 225
Belgacem, Ali 189
Benevides, Nathalia 163
Bey, Kadda Beghdad 189
Bicevska, Zane 197
Bicevskis, Janis 197
Bielecki, Włodzimierz 111
Biot, Francois 73
Buckley, Steve 73
Buryakova, Anastasia 45, 171
- C**archiolo, Vincenza 151
Correvon, Marc 73
Cunha, Douglas 225
Czarnul, Paweł 105
- D**anielewicz-Betz, Anna 37
Debicki, Olivier 73
Dörpinghaus, Jens 13
Dudnik, Gabriela 73
- E**inabadi, Behnam 61
- F**abre, Christian 73
Foucault, Julie 73
Franczyk, Bogdan 145
- G**haderi, Seyed Farid 61
Grandjean, Nathalie 73
Gruzina, Ulia 45, 171
Gyseghem, Jean-Marc Van 73
- H**attori, Yuto 3
Herveg, Jean 73
- J**ackson, Carl 73
Jacobs, Marc 13
Jaworski, Tomasz 217
Jeong, Minki 129
- K**ajiwara, Yusuke 3
Kim, Changick 129
Kim, Yoonhyung 129
Kiseleva, Anna 83
Kucharski, Jacek 217
Kucharski, Przemysław 217
- L**aszko, Łukasz 123
Lee, Jiwon 129
Lemos, Lucas 225
Lesecq, Suzanne 73
Lipka, Richard 233
Loria, Mark Phillip 151
Luzhnov, Petr 83
- M**aciel, Cristiano 163
Malgeri, Michele 151
Mareau, Nicolas 73
Mathewson, Alan 73
Matteo, Andrea di 73
McGibney, Alan 73
Memari, Pedram 61
Mozgovoy, Maxim 37
Murawski, Krzysztof 95
- N**acer, Hassina 189
Nam, Do-Won 129
Neder, Renato 163
Němec, Radek 179
Nikiforova, Anastasija 197
Nunes, Michel 225
- O**ditis, Ivo 197
O’Keeffe, Rosemary 73
O’Murchu, Cian 73
Ouvry, Laurent 73
- P**ąk, Karol 23
Pałkowski, Marek 111
Palma, Vincenza Di 73
Pinto, João Victor 225
Potuzak, Tomas 233
- Q**uaglia, Fabio 73

R abelo, Oliván	163	Tomaszuk, Dominik	27
Ramalho, Paulo	163	Trubnikov, Vladislav	45, 171
Razavi, Joseph	73	U lacha, Grzegorz	135
Rea, Susan	73	V arnavskiy, Andrew	45, 171
Rojas, David	73	Varone, Sacha	53
Romanowski, Andrzej	217	Villa, Giuseppe	73
Rot, Artur	45, 171	W awrzynczak, Anna	115
S chöne, Eric	145	Wernik, Cezary	135
Sebechenko, Ekaterina	45, 171	Wetering, Rogier Van de	87
Senczyna, Krzysztof	179	X u, Qiwen	241
Sevrin, Loic	73	Y amaguchi, Hiroshi	37
Shamaev, Dmitry	83	Yoo, Wonyoung	129
Shimakawa, Hiromitsu	3	Z ambra, Elizandra	163
Sielski, Dawid	217	Žentara, Tomasz	95
Skowron, Philipp	145	Zhu, Huibiao	241
Steinhage, Volker	13	Zhu, Longfei	241
Szelągowski, Marek	205		
T anaka, Tomoki	3		
Tarenskeen, Deborah	87		
Thiry, Florence	73		
Toja, Marco	151		