International Journal on Advances in Intelligent Systems

2018 vol. 11 nr. 1&2

The *International Journal on Advances in Intelligent Systems* is Published by IARIA.
ISSN: 1942-2679
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Intelligent Systems
Volume 11, Number 1 & 2, 2018

Kyu-Chul Lee, Chungnam National University, South Korea
Tracey Kah Mein Lee, Singapore Polytechnic, Republic of Singapore
Daniel Lemire, LICEF Research Center, Canada
Haim Levkowitz, University of Massachusetts Lowell, USA
Kuan-Ching Li, Providence University, Taiwan
Tsai-Yen Li, National Chengchi University, Taiwan
Yangmin Li, University of Macau, Macao SAR
Jian Liang, Nimbus Centre, Cork Institute of Technology, Ireland
Haibin Liu, China Aerospace Science and Technology Corporation, China
Lu Liu, University of Derby, UK
Qing Liu, The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Shih-Hsi "Alex" Liu, California State University - Fresno, USA
Xiaoqing (Frank) Liu, Missouri University of Science and Technology, USA
David Lizcano, Universidad a Distancia de Madrid, Spain
Henrique Lopes Cardoso, LIACC / Faculty of Engineering, University of Porto, Portugal
Sandra Lovrencic, University of Zagreb, Croatia
Jun Luo, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
Prabhat K. Mahanti, University of New Brunswick, Canada
Jacek Mandziuk, Warsaw University of Technology, Poland
Herwig Mannaert, University of Antwerp, Belgium
Yannis Manolopoulos, Aristotle University of Thessaloniki, Greece
Antonio Maria Rinaldi, Università di Napoli Federico II, Italy
Ali Masoudi-Nejad, University of Tehran, Iran
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Zulfiqar Ali Memon, Sukkur Institute of Business Administration, Pakistan
Andreas Merentitis, AGT Group (R&D) GmbH, Germany
Jose Merseguer, Universidad de Zaragoza, Spain
Frederic Migeon, IRIT/Toulouse University, France
Harald Milchrahm, Technical University Graz, Institute for Software Technology, Austria
Les Miller, Iowa State University, USA
Marius Minea, University POLITEHNICA of Bucharest, Romania
Yasser F. O. Mohammad, Assiut University, Egypt
Shahab Mokarizadeh, Royal Institute of Technology (KTH) - Stockholm, Sweden
Martin Molhanec, Czech Technical University in Prague, Czech Republic
Charalampos Moschopoulos, KU Leuven, Belgium
Mary Luz Mouronte López, Ericsson S.A., Spain
Henning Müller, University of Applied Sciences Western Switzerland - Sierre (HES SO), Switzerland
Susana Munoz Hernández, Universidad Politécnica de Madrid, Spain
Bela Mutschler, Hochschule Ravensburg-Weingarten, Germany
Deok Hee Nam, Wilberforce University, USA
Fazel Naghdy, University of Wollongong, Australia
Joan Navarro, Research Group in Distributed Systems (La Salle - Ramon Llull University), Spain
Rui Neves Madeira, Instituto Politécnico de Setúbal / Universidade Nova de Lisboa, Portugal
Andrzej Niesler, Institute of Business Informatics, Wroclaw University of Economics, Poland
Kouzou Ohara, Aoyama Gakuin University, Japan
Jonice Oliveira, Universidade Federal do Rio de Janeiro, Brazil
Ian Oliver, Nokia Location & Commerce, Finland / University of Brighton, UK
Michael Adeyeye Oluwasegun, University of Cape Town, South Africa
Sascha Opletal, University of Stuttgart, Germany
Fakri Othman, Cardiff Metropolitan University, UK
Enn Õunapuu, Tallinn University of Technology, Estonia
Jeffrey Junfeng Pan, Facebook Inc., USA
Hervé Panetto, University of Lorraine, France

Malgorzata Pankowska, University of Economics, Poland
Harris Papadopoulos, Frederick University, Cyprus
Laura Papaleo, ICT Department - Province of Genoa & University of Genoa, Italy
Agis Papantoniou, National Technical University of Athens, Greece
Thanasis G. Papaioannou, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Andreas Papasalouros, University of the Aegean, Greece
Eric Paquet, National Research Council / University of Ottawa, Canada
Kunal Patel, Ingenuity Systems, USA
Carlos Pedrinaci, Knowledge Media Institute, The Open University, UK
Yoseba Penya, University of Deusto - DeustoTech (Basque Country), Spain
Cathryn Peoples, Queen Mary University of London, UK
Asier Perallos, University of Deusto, Spain
Christian Percebois, Université Paul Sabatier - IRIT, France
Andrea Perego, European Commission, Joint Research Centre, Italy
Mark Perry, University of Western Ontario/Faculty of Law/ Faculty of Science - London, Canada
Willy Picard, Poznań University of Economics, Poland
Agostino Poggi, Università degli Studi di Parma, Italy
R. Ponnusamy, Madha Engineering College-Anna University, India
Wendy Powley, Queen's University, Canada
Jerzy Prekurat, Canadian Bank Note Co. Ltd., Canada
Didier Puzenat, Université des Antilles et de la Guyane, France
Sita Ramakrishnan, Monash University, Australia
Elmano Ramalho Cavalcanti, Federal University of Campina Grande, Brazil
Juwel Rana, Luleå University of Technology, Sweden
Martin Randles, School of Computing and Mathematical Sciences, Liverpool John Moores University, UK
Christoph Rasche, University of Paderborn, Germany
Ann Reddipogu, ManyWorlds UK Ltd, UK
Ramana Reddy, West Virginia University, USA
René Reiners, Fraunhofer FIT - Sankt Augustin, Germany
Paolo Remagnino, Kingston University - Surrey, UK
Sebastian Rieger, University of Applied Sciences Fulda, Germany
Andreas Riener, Johannes Kepler University Linz, Austria
Ivan Rodero, NSF Center for Autonomic Computing, Rutgers University - Piscataway, USA
Alejandro Rodríguez González, University Carlos III of Madrid, Spain
Paolo Romano, INESC-ID Lisbon, Portugal
Agostinho Rosa, Instituto de Sistemas e Robótica, Portugal
José Rouillard, University of Lille, France
Paweł Różycki, University of Information Technology and Management (UITM) in Rzeszów, Poland
Igor Ruiz-Agundez, DeustoTech, University of Deusto, Spain
Michele Ruta, Politecnico di Bari, Italy
Melike Sah, Trinity College Dublin, Ireland
Francesc Saigi Rubió, Universitat Oberta de Catalunya, Spain
Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Yacine Sam, Université François-Rabelais Tours, France
Ismael Sanz, Universitat Jaume I, Spain
Ricardo Sanz, Universidad Politecnica de Madrid, Spain
Marcello Sarini, Università degli Studi Milano-Bicocca - Milano, Italy
Munehiko Sasajima, I.S.I.R., Osaka University, Japan
Minoru Sasaki, Ibaraki University, Japan
Hiroyuki Sato, University of Tokyo, Japan
Jürgen Sauer, Universität Oldenburg, Germany
Patrick Sayd, CEA List, France
Dominique Scapin, INRIA - Le Chesnay, France

Kenneth Scerri, University of Malta, Malta
Rainer Schmidt, Austrian Institute of Technology, Austria
Bruno Schulze, National Laboratory for Scientific Computing - LNCC, Brazil
Ingo Schwab, University of Applied Sciences Karlsruhe, Germany
Wieland Schwinger, Johannes Kepler University Linz, Austria
Hans-Werner Sehring, Namics AG, Germany
Paulo Jorge Sequeira Gonçalves, Polytechnic Institute of Castelo Branco, Portugal
Kewei Sha, Oklahoma City University, USA
Roman Y. Shtykh, Rakuten, Inc., Japan
Robin JS Sloan, University of Abertay Dundee, UK
Vasco N. G. J. Soares, Instituto de Telecomunicações / University of Beira Interior / Polytechnic Institute of Castelo Branco, Portugal
Don Sofge, Naval Research Laboratory, USA
Christoph Sondermann-Woelke, Universitaet Paderborn, Germany
George Spanoudakis, City University London, UK
Vladimir Stantchev, SRH University Berlin, Germany
Cristian Stanciu, University Politehnica of Bucharest, Romania
Claudius Stern, University of Paderborn, Germany
Mari Carmen Suárez-Figueroa, Universidad Politécnica de Madrid (UPM), Spain
Kåre Synnes, Luleå University of Technology, Sweden
Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland
Yehia Taher, ERISS - Tilburg University, The Netherlands
Yutaka Takahashi, Senshu University, Japan
Dan Tamir, Texas State University, USA
Jinhui Tang, Nanjing University of Science and Technology, P.R. China
Yi Tang, Chinese Academy of Sciences, China
John Terzakis, Intel, USA
Sotirios Terzis, University of Strathclyde, UK
Vagan Terziyan, University of Jyvaskyla, Finland
Lucio Tommaso De Paolis, Department of Innovation Engineering - University of Salento, Italy
Davide Tosi, Università degli Studi dell'Insubria, Italy
Raquel Trillo Lado, University of Zaragoza, Spain
Tuan Anh Trinh, Budapest University of Technology and Economics, Hungary
Simon Tsang, Applied Communication Sciences, USA
Theodore Tsiligiridis, Agricultural University of Athens, Greece
Antonios Tsourdos, Cranfield University, UK
José Valente de Oliveira, University of Algarve, Portugal
Eugen Volk, University of Stuttgart, Germany
Mihaela Vranić, University of Zagreb, Croatia
Chieh-Yih Wan, Intel Labs, Intel Corporation, USA
Jue Wang, Washington University in St. Louis, USA
Shenghui Wang, OCLC Leiden, The Netherlands
Zhonglei Wang, Karlsruhe Institute of Technology (KIT), Germany
Laurent Wendling, University Descartes (Paris 5), France
Maarten Weyn, University of Antwerp, Belgium
Nancy Wiegand, University of Wisconsin-Madison, USA
Alexander Wijesinha, Towson University, USA
Eric B. Wolf, US Geological Survey, Center for Excellence in GIScience, USA
Ouri Wolfson, University of Illinois at Chicago, USA
Yingcai Xiao, The University of Akron, USA
Reuven Yagel, The Jerusalem College of Engineering, Israel
Fan Yang, Nuance Communications, Inc., USA
Zhenzhen Ye, Systems & Technology Group, IBM, US A

International Journal on Advances in Intelligent Systems
Volume 11, Numbers 1 & 2, 2018

CONTENTS

pages: 1 - 12
Youth with ME/CFS: Untangling Complexities in Technology Design to Support Everyday Living
Jorun Børsting, University of Oslo, Norway
Alma Leora Culén, University of Oslo, Norway

pages: 13 - 24
Architectural Backbone Evaluation for Data Stream Processing within the WINNER DataLab - A Project Focused Point of View
Sebastian Apel, Chair for Software Engineering, Friedrich-Schiller-University Jena, Germany
Florian Hertrampf, Chair for Software Engineering, Friedrich-Schiller-University Jena, Germany
Steffen Späthe, Chair for Software Engineering, Friedrich-Schiller-University Jena, Germany

pages: 25 - 34
Adaptive Portfolio Asset Allocation Optimization with Deep Learning
Samer Obeidat, Investifai Inc., Canada
Daniel Shapiro, School of Electrical Engineering and Computer Science, University of Ottawa; Lemay Solutions Consulting Inc.; Investifai Inc., Canada
Mathieu Lemay, Lemay Solutions Consulting Inc., Canada
Mary Kate MacPherson, Lemay Solutions Consulting Inc., Canada
Miodrag Bolic, School of Electrical Engineering and Computer Science, University of Ottawa, Canada

pages: 35 - 45
Cartesian Systemic Pulsation – A Model for Evolutive Improvement of Incomplete Symbiotic Recursive Systems
Marta Franova, LRI, UMR8623 du CNRS & INRIA, France
Yves Kodratoff, LRI, UMR8623 du CNRS & INRIA, France

pages: 46 - 54
From Compartments to Agents via Fuzzy Models - Modeling and Analysis of Complex Behavior of Physiological Systems
Sebastian Meszyński, Faculty of Physics, Astronomy, and Informatics Nicolaus Copernicus University, Poland
Oleksandr Sokolov, Faculty of Physics, Astronomy, and Informatics Nicolaus Copernicus University, Poland
Aleksandra Mreła, Faculty of Technology, Kujawy and Pomorze University in Bydgoszcz, Poland

pages: 55 - 67
Study of the Appropriation of Groupware in the Context of Remote Collaborative Design. Aspectuality via time, occurrence and changes
Samia Ben Rajeb, Université Libre de Bruxelles, Belgium
Shima Shirkhodaei, University of Liege, Belgium
Pierre Leclercq, University of Liege, Belgium

pages: 68 - 80
User-guided Graph Exploration: A Framework for Algorithmic Complexity Reduction in Large Data Sets
Tim Grube, Telecooperation Group, TU Darmstadt, Germany
Florian Volk, House of IT, Germany
Max Mühlhäuser, Telecooperation Group, TU Darmstadt, Germany

Suhas Bhairav, iTrust, Singapore University of Technology and Design, Singapore
Vinay Sachidananda, iTrust, Singapore University of Technology and Design, Singapore
Yuval Elovici, iTrust, Singapore University of Technology and Design, Singapore

pages: 81 - 90
Reporting and Analyzing Student Behavior in 3D Virtual Worlds
Aliane Loureiro Krassmann, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil
Felipe Becker Nunes, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil
Tito Armando Rossi Filho, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil
Liane Margarida Rockenbach Tarouco, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil
Magda Bercht, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

pages: 91 - 104
Reconstruction and Web-based Editing of 3D Objects from Photo and Video Footage for Ambient Learning Spaces
David Bouck-Standen, Institute for Multimedia and Interactive Systems, University of Luebeck, Germany
Alexander Ohlei, Institute for Multimedia and Interactive Systems, University of Luebeck, Germany
Sven Höffler, Institute for Multimedia and Interactive Systems, University of Luebeck, Germany
Viktor Daibert, Institute for Multimedia and Interactive Systems, University of Luebeck, Germany
Thomas Winkler, Institute for Multimedia and Interactive Systems, University of Luebeck, Germany
Michael Herczeg, Institute for Multimedia and Interactive Systems, University of Luebeck, Germany

pages: 105 - 114
An Introduction to Edge Computing and A Real-Time Capable Server Architecture
Volkan Gezer, DFKI, Germany
Jumyung Um, DFKI, Germany
Martin Ruskowski, DFKI, Germany

pages: 115 - 126
Deduction System for Decision Logic Based on Many-valued Logics
Yotaro Nakayama, Nihon Unisys, Ltd., Japan
Seiki Akama, C-Republic, Inc., Japan
Tetsuya Murai, Chitose Institute of Science and Technology, Japan

pages: 127 - 133
Pro2Screen -- High-Fidelity-Prototyping of Mobile Enterprise Applications using Process Models
Leo Fuhr, RheinMain University of Applied Sciences, Germany
Matthias Jurisch, RheinMain University of Applied Sciences, Germany
Stephan Böhm, RheinMain University of Applied Sciences, Germany
Tobias Braumann, RheinMain University of Applied Sciences, Germany
Bodo Igler, RheinMain University of Applied Sciences, Germany

Youth with ME/CFS: Untangling Complexities in Technology Design to Support Everyday Living

Jorun Børsting
Department of Informatic
University of Oslo
Oslo, Norway
jorubo@ifi.uio.no

Alma L. Culén
Department of Informatics
University of Oslo
Oslo, Norway
almira@ifi.uio.no

Abstract—**In this paper, we argue that the technology design needs to take a more holistic perspective, beyond opportunities offered by the technology development alone. This is especially important when developing technologies for complex contexts, such as healthcare, and for users who are at risk. Furthermore, we present a sense-making process, informed by tools and methods from social science, complex systems design and design thinking, and intended for human-computer interaction designers and researchers to broaden their understanding of entanglement of technology, users, and the design context. This is particularly important when working in a new application area, where one is not an expert. We illustrate our approach by exploring if and how digital tools could support youth with Myalgic Encephalomyelitis (ME), also known as a Chronic Fatigue Syndrome (CFS), to better cope with everyday challenges. Informed by previous research related to our field of inquiry, we combine ethnographic research, Actor-Network Theory adapted for design, and Giga-mapping (a complex systems design technique). We find that this combination of methods and techniques supports deeper reflections over the domain complexities, and makes it easier to avoid designing technologies for adolescents with ME/CFS that do not offer real support in day-to-day living.**

Keywords-complex systems; sense-making; ME/CFS; ANT; healthcare; design thinking.

I. INTRODUCTION

This paper extends our previous work [1], presented at the eTELEMED 2016 (International Conference on eHealth, Telemedicine, and Social Medicine), on sense-making when designing new technologies for adolescents with ME/CFS.

Increasing attention is given to diverse self-management applications and technologies that provide support to people with health challenges to improve or maintain the quality of everyday living, sustain active recovery processes or monitor relevant health data. However, the knowledge required to develop such applications successfully is often multi-disciplinary. Thus, human-computer interaction (HCI) researchers and designers need to reach some level of insight into these other disciplinary perspectives to gain a holistic view of new design contexts, beyond the focus on technology alone.

Healthcare related problems can often be labeled as wicked [2]–[4]. Wicked problems have been discussed by Rittel and Webber already in 1973 [5] but gained additional visibility and attention recently through design-driven approaches such as design thinking [6]. Wicked problems are incomplete, contradictory, changing and interdependent. Working with them is difficult since making changes in one area (e.g., making an app for monitoring health data) may affect other areas (e.g., how and with whom to share information when data shows risks), potentially causing even more challenges and making visible previously unidentified issues. Thus, isolating a problem and proceeding to find a solution for it is not the best way of dealing with problems in complex domains. Rather, untangling complexities and gaining a deeper understanding of the context, gives a possibility to identify underlying issues, patterns and good opportunities for a design that may lead to mitigation of the challenges that users, and others involved, meet (such as medical professional, patient organizations, parents, or teachers). Thus, the design of healthcare technologies that support users and are well situated in the use context requires a broader and deeper inquiry into the nature of problems and their relation to the situated design space [7].

We describe our approach and discuss the initial sense-making in complex domains, a process that enables this broader and deeper understanding of the context, and design options. Our approach is illustrated using the inquiry into the possibility of designing supportive technologies for adolescents with Myalgic Encephalomyelitis (ME), often referred to as the Chronic Fatigue Syndrome (CFS), or ME/CFS. Almost no research results were to be found on the design of new technologies made specifically for this user group. Without understanding challenges that these adolescents meet in their day-to-day life, it is not possible to envision technological solution(s), be it self-management tools, or other systems or services that could be useful in their everyday lives. We, thus, propose a framework for sense-making and finding opportunities to resolve issues that help untangle some of the complexity within a design context.

In recent years, the increase of chronic diseases, in particular among youth, has been significant [8]. Among these diseases, ME/CFS has become a growing concern, not only for those suffering from it and their families, but also for the medical science, governmental health management, and society at large. Tweens and adolescents who are affected by the illness would be, if healthy, considered to be willing adopters and users of technology [9]. However, even for their healthy peers, intensive use of technology may lead to negative health consequences [10]. Thus, for those with ME/CFS, it is crucial that technology does not introduce a new set of problems or worsens their health condition [11].

From this perspective, we consider youth with ME/CFS to be a group of vulnerable users [12]. The vulnerability is to be understood as a set of risks that these adolescents are exposed to. Understanding these risks, and how to reduce or eliminate them, is a step towards understanding design and technology design spaces when considering designing for and with this user group.

Furthermore, the understanding of the design space in complex contexts may require an understanding of several different fields pertinent to the context. Understandings can be at the micro-level (e.g., how a specific solution affects the user) and understandings at the macro-level (such as social, cultural and governmental influences on the design space). We have chosen Actor-Network Theory (ANT) as a tool that can help produce rich descriptions of the interrelated social and material processes at different levels [13][14], Giga-mapping to represent and communicate our thinking about design, and finally, a firm focus on users and use. ANT is borrowed from Science and Technology Studies (STS) and Giga-mapping from System Oriented Design (S.O.D). In combination, for HCI researchers and designers, they offer a new tool for initial sense-making. We find that the approach helps develop and communicate insights efficiently, increases awareness of a broad range of issues and points to their interconnectedness. Finally, it engages HCI researchers in innovative and exploratory practices within an unfamiliar design context and enables critical reflection on own position and values. As a consequence, the process points in a more informed manner towards areas where prototyping is helpful, and when finishing prototypes to the extent that they can be used in real life use explorations is essential [15][16][17]. In this work, however, we focus only on how we gained insights and not on any specific solutions.

The paper is structured as follows: the contextual background is covered in Section II, including relevant research findings on ME/CFS and technology design, and describing the methodology, methods, and tools (design thinking, ANT, and Giga-mapping). Our first expert session is described in Section III and analyzed, using ANT, in Section IV. Section V addresses the combined approach of ANT and Giga-mapping. Section VI describes the Giga-mapping workshop. The second expert session is presented in VII. Section VIII showcases the first realized technology design. The conclusion and the future work follow in Section IX.

II. Background

Previously published research work is a usual starting place when exploring a new application domain. In the case addressed in this paper, some understanding from several disciplines is needed. First and foremost, design for adolescents, in particular, those with health challenges, is often outside the main research focus within the HCI. Designing for children is a well-established subfield of HCI, with developed tools, methods and methodologies. The work with adolescents has not attracted the same kind of attention. In the case presented in this paper, adolescents are exposed to numerous health and other risks and can be seen as a vulnerable user group. To understand risks that they are exposed to, grasping what it means to be an adolescent living with ME/CFS is crucial. Empathy [18], interviews, observations, and other methods may be used to gain insight into lives of adolescents with the illness. However, some familiarity with basic scientific results from current and published medical research on the condition is also necessary, even though it is outside of our main field of competence. The medical research results helped to identify a broad set of risks that adolescents with ME/CFS face, including isolation and stigma.

We divide this section into four parts A) medical understanding of the condition and associated risks, B) review of previous research, focusing on technology for ME/CFS, C) design thinking and vulnerable users, D) pragmatic use of ANT and mapping in design.

A. A medical understanding of adult and pediatric ME/CFS and its effects: the challenging nature of the disease

ME/CFS ME/CFS is a debilitating multisystem illness resulting in a plethora of symptoms that include severe physical and cognitive exhaustion, confusion, difficulties with memory, concentration, sensitivity to light and noise among others, see [19], [20]. What causes the illness remains an enigma, and the condition is presently researched on a broad scale, e.g., [21]. Many theories, ranging from viral infections to psychological stress have been proposed [19]. Some studies suggest that the immune system may be chronically active in people with CFS. This might relate to a theory that ME/CFS is caused by an abnormal reaction to common infectious agents. This theory could link ME/CFS to autoimmune diseases such as the Lyme disease and the Epstein-Barr virus.

A review of thirty-four qualitative studies on ME/CFS, see [22], was done and the authors find that “For sufferers, illness development influenced identity, reductions in functioning, and coping. Physician-specific themes described lack of awareness about ME/CFS and recommended improvement in educational resources. Intersecting themes expressed issues with diagnosis creating tensions and fueling the stigmatization of ME/CFS”. Also, some research points in the direction where people with ME/CFS themselves help to perpetuate their condition. For example, Afari and Buchwald state that “current knowledge about chronic fatigue syndrome suggests that genetic, physiological, and psychological factors work together to predispose an individual to the condition and to precipitate and perpetuate the illness” and, further, that “sufferers’ perceptions, illness attributions, and coping skills may help to perpetuate the illness” [19, p. 230]. Such results, unless verified carefully, may lead to and perpetuate stigmatization. Other research, such as that of Geelen et al. [23], or Winger et al. [24], offers insight into personality issues relevant for understanding the youth with ME/CFS.

The treatment of those diagnosed with ME/CFS is highly individualized, frequently symptom-based, and includes both pharmacological [21] and behavioral approaches [25]. The most common form of help to those with illness in Norway [26] are self-management courses, offering guidance on how to stabilize the symptoms, find a balance between rest and activity, and adjust to living a life with ME. This is in line with the trend to promote self-management and to, wherever appropriate, foster patient participation in care, and especially so for people with chronic diseases [27]. Self-management often involves the management of medical conditions, behavior, or emotions [28], [29]. For users with ME/CFS, self-management also includes the management of so-called energy modulation. A person with ME/CFS is considered to have a limited amount of energy available for use per day. If an activity takes too much energy, the overuse results in extra exhaustion over a period of time lasting from several hours to several days, leading to increased pain levels, sensitivity and overall aggravate the physical and cognitive condition. The problem is often that a person with ME/CFS do not know what amount of energy is required for an ordinary activity such as, for example, meeting a friend for a cup of coffee. Therefore, the energy modulation management is difficult.

A user study carried out by the Norwegian ME/CFS association [20] found that 40% of people with ME/ CFS did not receive medical help with relieving symptoms caused by the illness. When asked what kind of help they would like to receive from the healthcare sector, they indicated professional assistance with sleep problems, stress management, stomach problems and the general pain relief.

For children and youth with ME/CFS, the situation is additionally difficult because it is harder to make their voice on the matter heard:

"Do children and adolescents suffer from ME/CFS? Simple common sense tells most parents, teachers and doctors that they do and often more severely than adults. Yet there remain a sizeable proportion of professionals in Health Care, Education and Social Services who are still prepared to ascribe the numerous, disabling but seemingly unconnected symptoms of this illness in young people to anorexia, depression, school phobia or dysfunctional family background. All are, at least, agreed that the illness presents a considerable economic, educational and social problem" [30].

Resonating with these findings, and in part, because the illness does not leave visible marks on a person, people with ME/CFS are often dismissed as suffering from psychological problems, or worse, stigmatized by general population [22].

Those diagnosed with ME/CFS have to bear the medical uncertainty of what causes the disease and, as a consequence, the absence of an effective diagnostic test and treatment for the illness [31]. Furthermore, social and political uncertainties and risks are tied to the lack of medical understandings.

The youth is affected by nearly the same rates as adults, from about the age of eleven [32]. For someone to be diagnosed with ME/CFS, the symptoms need to be present to such a degree that they clearly limit person's ability to carry out ordinary daily activities [33]. Youth are classified into four groups: mild (approximately 50% reduction in pre-illness activity level), moderate (mostly housebound), severe (mostly bedridden), or very severe (totally bedridden and in need of help with basic functions) [34].

B. Previous research on ME/CFS in adolescence and technology

As mentioned in the introduction, there is very little in the literature on supporting and assisting people with ME/CFS, including adolescents, using technology. People with ME/CFS, with their multiple challenges at the physical and cognitive level, coupled with reduced tolerance to light and noise, may not be the prime candidates for the use of screens for receiving information, whether the screens are on tablets, smartphones, or personal computers. Yet, all solutions that we could find in the previous research were screen based. While such solutions may be a good option for healthy adolescents, they have limitations for those with ME/CFS. Thus, before deciding on any new technological solutions for adolescents with ME/CFS, one needs to make sure that proposed solutions do not introduce new problems and health risks for their users.

Previous research shows that social isolation and access to education are important challenges facing these adolescents. Several researchers explored solutions in these two directions, studying either the social media as means of reducing isolation, or platforms for education. General use of social media in medical care was described in [35], where authors have analyzed and synthesized 76 articles, 44 websites, and 11 policies/reports and presented findings according to 10 different categories of social media such as blogs, microblogs, social networks, professional networks, thematic networks, wikis, sharing sites and others (represented by the Second Life). Findings, in particular, those related to the Second Life are relevant for adolescents with ME/CFS and have been taken further in the work of Best and Butler [36], [37]. The paper [37] describes how a virtual support center was constructed in the Second Life, featuring meeting areas, relaxation areas, library resources and a gallery of art by and for people with ME/CFS. However, the results of the investigation gave mixed conclusions, possibly reflecting the fact that certain level of mastery of the virtual environment was needed, and that was not possible for all users to accomplish. Also, the physical and cognitive condition of people with ME/CFS may affect their willingness to dedicate energy to learning about new virtual environments. The Second Life application provided information related to the condition and was aiming to address loneliness and social isolation.

Considering the educational platforms for those with ME/CFS, the paper on e-learning, [38], can be pointed out as relevant. In this study, e-learning platforms were explored as an opportunity for children with ME/CFS to participate in classroom activities even though they are staying at home. As the authors state, this is especially relevant to Scotland, where many live far away from schools. However, the access to school and educational materials is only the first hurdle in receiving regular education and does not address other problems that youth with ME/CFS often have.

C. Design Thinking and Vulnerable users?

Design Thinking may be defined in many different ways. For example, it may be defined as a process that fosters innovative and creative thinking [39], or as an approach to mitigate complex problems through design [6], [40]. Design thinking is also often described as a user or human-centered process, resting on three pillars: empathy with the user, rapid prototyping and abductive thinking [39]. In [41], the authors engage in an important discussion on the role of design thinking in research and designerly practice, suggesting core concepts that include reflexive practice, meaning-making and designerly ways of knowing.

We argue that in complex design contexts, sense-making is also essential. It is to be distinguished from meaning-making, which has to do with design research, practice and designerly ways of knowing. Sense-making, as we use it, applies to the understanding of the context. It requires careful framing and setting of contextual boundaries, balancing the broadness and the depth of inquiries. Giga-mapping, [42], can be viewed as a tool for sense-making. As the name suggests, it allows for mapping and visualization of a variety of perspectives and concerns. The making of a Giga-map facilitates thinking and communication. By its large physical size, it invites participation and collective production of understandings regarding the design context. Furthermore, the context can be mapped out in layers (e.g., understandings on micro, mezzo, and macro levels), fostering further insight into inter-relatedness of problems. As discussed later in this paper, maps, including Giga-maps, are important actors in design networks.

When a design context has users at risk in centrum, the term ‘vulnerable users’ is often used. In [11], authors argue that the term ‘vulnerable’ should not be seen as stigmatizing. Rather, it implies a set of risks that users are exposed to. When identified, these risks are helpful in defining design goals that aim to reduce or eliminate them. They further argue “the awareness of risks/vulnerabilities in a design situation may be helpful in designing better products for vulnerable people” [11, p. 3]. Thus, in design for and with vulnerable users, risks need to be recognized and mapped out already in the initial, sense-making processes, prior to engagement with any concrete design ideas. For example, adolescents suffering from ME/CFS often have limited ability to read from digital devices, or even just look at a smartphone or computer screen (due to the sensitivity to light). Thus, they may be at risk of not being able to use any screen-based solutions because of their illness symptoms.

When working with vulnerable users, it is important to take into consideration ethical concerns as part of the reflection and meaning-making. Ideally, both the sense-making and the design processes should include real users, in co-design setting. With vulnerable users, this may not always be possible. Those speaking on their behalf must be able to represent them well, i.e., need to be able to convey their interests, challenges, and experiences.

D. Pragmatic use of Actor-Network Theory and mapping

We propose pragmatic use of ANT to inspire analysis of design opportunities in the initial, sense-making phases of a design process. This aims at creating a holistic picture around adolescents with ME/CFS, communicated and represented by for example, a Giga-map. The process helps to explore opportunities for technology to offer support in their everyday lives. It also helps to imagine and discuss consequences of future technologies before they are made.

ANT is traditionally deployed as a framework to gain knowledge about heterogeneous networks, consisting of diverse human and non-human actors. The inclusion of non-human actors differentiates ANT from other social theories employed in information systems research. ANT focuses on both the social and the technical factors involved in the research case; all the social-technical elements are incorporated into networks of actors/actants. Therefore, ANT can also be used to theorize about digital artefacts. Latour describes the concept of networks: “Behind the actors, others appear; behind one set of intentions there are others; between the (variable) goals and the (variable) desires, intermediate goals and implications proliferate, and they all demand to be taken into account” [44, p. 100]. The social-technical focus implies analyzing both the visible (the technical/objects) and the invisible (the social) and then distinguishing the relationships between them. Those relationships can be both semiotic and material simultaneously. Combined, they constitute a network of actors that acts as a whole [45]. Tatnall and Gilding (2005) argue that ANT can be particularly useful for studies in areas that involve consideration of political and social issues in information systems. They further mention interface design, usability testing, and the use of distributed systems within organizations as examples of possible application areas [46].

Currently, there is an increasing focus on how ANT could be fruitful in the collaborative design of digital artefacts [14], [47], [48]. Stuedahl and Smørdal state that “involving ANT concepts in co-design does help to frame co-design processes within the wider context and consequences of emerging knowledge development” [49, p. 204]. A number of other researchers focused on ANT and design. For example, Schoffelen, in [50], is concerned with designing visualizations that make things public, to interest and engage people in participatory processes. Storni, in [14], explored how ANT can be used to offer an alternative perspective to co-designers, focusing on “a convention from Latour’s call for risky descriptions to a call to design things together” [14, p.167]. He further argues that ANT suggests three general turns to rethinking co-design and participatory design practices. The first turn addresses the question of what to design? It also involves the idea to first design actor networks, and then look for ways to map them. He further describes the resulting maps as not only descriptive but also as supporting participation in the design process. The second turn concerns the question of how to co-design and suggests the idea of designing as actor-networking (in public). The last turn, described as epistemological, involves what Storni describes as "moving from the idea of the designer as a network prince to the idea of the designer as an agnostic Prometheus" [14, p. 167]. Storni suggests that making maps is a way towards holistic and democratic design. The maps are themselves actors in the network, drawing things together through visual problematization, and equally importantly, offering everyone engaged in the design process to add, transform, remove, critique or highlight its elements.

III. THE FIRST EXPERT SESSION

As mentioned in the introduction, this study is informed by previous research on ME/CFS, and it uses a combination of ethnographic methods, research results, and analysis inspired by ANT and Giga-mapping to tease out real problems and design opportunities related to assistive technologies for young people with ME/CFS.

Regarding user research, we have organized a session with experts on ME/CFS. These experts have an in-depth knowledge of ME/CFS, also personally, or through their immediate family. They are in daily contact with children, youth, and adults with ME/CFS, and their families. Therefore, they have a broad, professional and personal knowledge of what it is to live with ME/CFS. These people came from 1) the Norwegian ME organization [26], 2) an organization that provides support to families of those with ME/CFS, 3) ME/CFS youth organization and 4) a medical doctor. The session lasted for close to three hours. A short part of the meeting was dedicated to getting to know what they do at the organizational level. The insights gained are incorporated in Fig. 1 and 2.

The session was also used to discuss future technologies that could provide some support in everyday lives for those with ME/CFS. Given the aforementioned deep knowledge the group shared, many design opportunities for future technologies, as well as novel uses of existing technology, were considered. Some of the suggestions that do not exist today, were brought up. We present some opportunities for design below.

The first opportunity was related to education, an important area for adolescents. It was an online e-learning solution with streamed lectures, accessible anytime. It was brought up that most young people with ME/CFS want to follow up on their studies, but they have to be able to take a break when needed. This addresses the need for a flexible educational system.

The second one was an App designed with a critical design approach with a specific focus on how the energy levels are different for people with ME/CFS. The App could simulate the amount of energy needed for simple everyday tasks to reduce prejudice and increase general understanding of the illness.

The third suggestion was another App that could be designed for self-management of the illness. The App could provide reminders to rest. It could also give short advice on things that have worked for others, which they could try to reduce illness symptoms. The App could also measure sound levels and give warnings when the levels are too high. Since exposure to high sound levels is known to increase illness symptoms, the App could also help visualize this, by showing sound measurements and giving reminders to use earplugs in spaces with high sound levels, e.g., when using public transportation. General personalized everyday reminders were also seen as valuable, as an increase in disease symptoms may lead to memory problems.

The fourth suggestion was an electronic diary that could be used to register symptoms daily. The knowledge and awareness gained through daily activities and their effects on illness symptoms were seen as valuable. This diary could also be a tool to explain better the current health status to healthcare professionals and social workers. For this diary to be used in periods with increased illness symptoms, it would be of utmost importance that the self-rating process is simple to conduct.

The fifth suggestion was related to wearable devices, such as the Fitbit. People with ME/CFS experienced the possibility of tracking activity, sleep patterns and energy expenditure as desirable. It was suggested that similar wearable technology, tailored to the needs of people with ME/CFS, could be valuable.

The sixth suggestion is a Social technology specially designed for people with ME/CFS, to meet and seek support from other people with the illness. It was expressed that it was much easier for people with ME, in most social settings to connect in a meaningful way with other people that also have the illness. Within the ME community there is a strong sense of community, but for adolescents, new ways to connect with peers that also have the illness could be helpful.

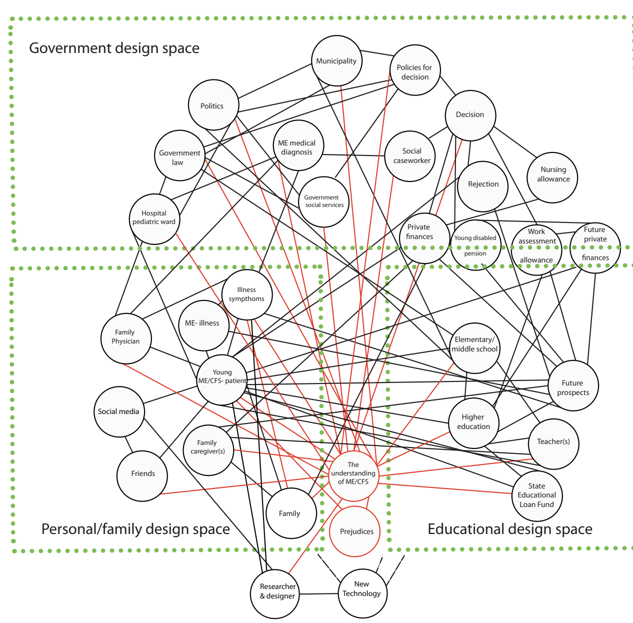


Figure 1. The first iteration of combined Giga-mapping/ANT.

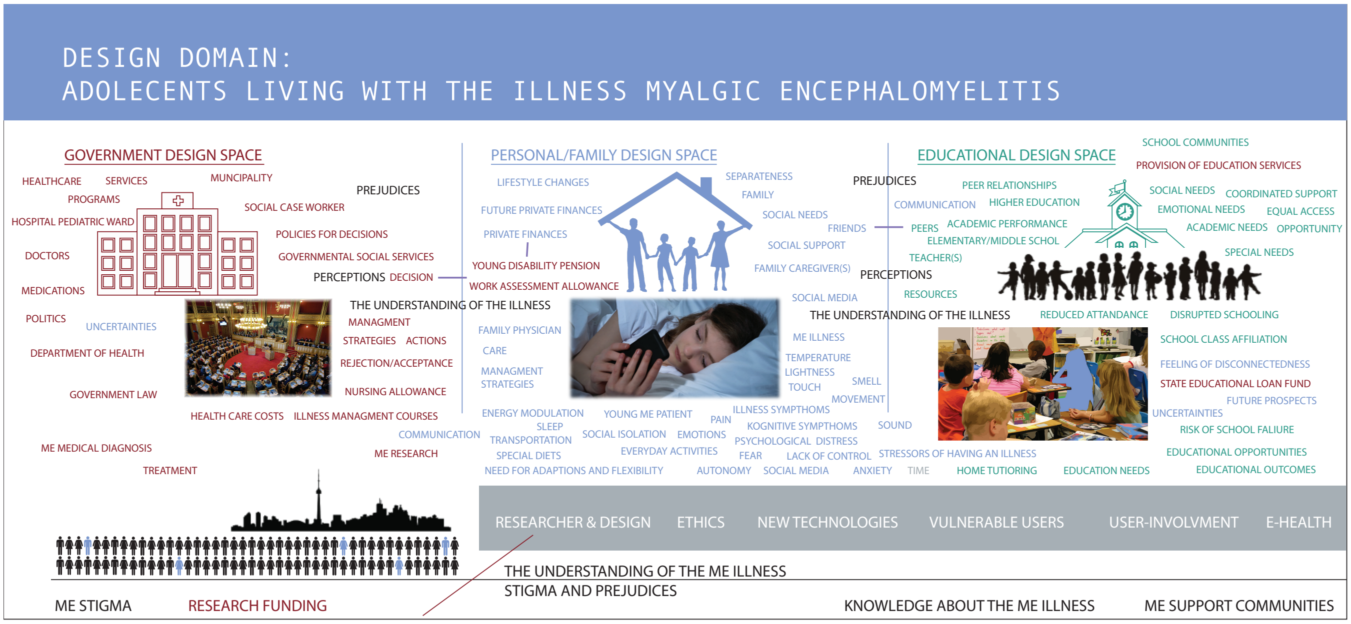


Figure 2. The new visualization of ANT/Giga-map.

IV. ANT Analysis of the Expert Session

Instead of considering any one of the mentioned suggestions as an automatic lead toward a design, and in isolation from other factors that influence people living with ME/CFS, we performed an extensive analysis, focusing on all of our data collected from the session with experts. The identified actors and issues from the expert session were also compared to findings from the relevant research literature. At first, the traditional ANT analysis was applied, but then, a more pragmatic, design and technology-oriented approach was deployed, in line with papers such as [14] and [48]. The result is shown in Fig. 2. The image is a visual representation, a Giga-map, at the macro level. All the relations shown are necessary for the initial sense-making process and following technology introduction. Each relation explicitly or implicitly contributes to a discussion around technology that could, in some way, aid adolescents with ME/CFS. Three large associated areas of the network emerge, related to what we named Personal/Family, Educational and Governmental design spaces. At present, those spaces are not strongly connected, and connections within each one of them are sparse. As mentioned earlier, also in the literature, there were very few results from HCI research within the research domain found.

In Fig. 1, a circle represents an actor and lines represent relations between actors. The number of actors shown is limited to the actors identified in the data analysis from the sense-making session with experts on ME/CFS. From an ANT perspective, all the relations presented in Fig. 1 is relevant to the development of the new technology. The actors in the case of research are characterized by their differences that affect the network. Actors in the network are both human (such as the ME/CFS sufferer, family members, hospital doctors and school teachers) and non-human (such as politics, prejudice, economics, and social media).This is called *symmetry* in ANT, where the non-human actors are integrated in the network in the same way as human actors. They all have a role in and influence on the network. These roles and influences are not predefined but continuously evolve. For example, a family member may consider himself or herself to be without prejudice in other situations, but in the network representing the youth with ME/CFS, they still can think that the adolescent is lazy rather than sick, and so contribute to maintaining the prejudice around the disease.

A concept in ANT is *material agency*. *Agency* can be seen as the actors 'capacity for influence' or the 'capacity for action' within the network [51]. As mentioned, all the different enrolled actors have a role and some influence in the network. *Agency* can be important to consider in early sense-making processes. Firstly, for the new technology design to be successfully adopted by the user group. The technology has to have some level of agency to not become obsolete within the design space. Secondly, because the technology introduction and use over time will potentially influence the agency of other actors in the network. This can be especially important with a vulnerable user group were unintended harmful effects due to such changes in agency can be particularly damaging.

The differences in patients' illness symptoms will give different results for the individual affected by ME/CFS. The level of illness impact on personal health is highly fluctuating not only on a day-to-day basis but also from person to person. The patients and different people interacting with the patient are also heterogeneous actors. They bring their different personalities, competencies, and experiences into the network. The knowledge of how these cases are currently is crucial for successfully identifying a technological design opportunity and this knowledge is gained through the relation with the people with ME/CFS.

Adolescents, various people that they interact with, as well as non-human entities also form several different heterogeneous networks. This directly relates to the wickedness of problems related to defining technologies that aim to help youth to have a better life, given their illness. Not only do adolescents suffer from the condition to varying degrees, but their health changes from day to day and period-wise as well. This changeability needs to be addressed as part of the sense-making.

Furthermore, the analysis revealed two mediators. Mediators are entities that increase difference among actors and, thus, should be carefully studied [52]. The mediators that we have distinguished impact how our society currently faces people with ME/CFS. In Fig. 1, the actors in the network, the two mediators are marked with red circles. They are the Understanding of ME/CFS and Prejudice. The Understanding of ME/CFS is seen as a contributing factor that defines relations between those with ME/CFS and many other actors in the network. It influences, for example, how the social service caseworkers or physicians handle youth with ME/CFS. The Prejudice seeks to impose its version of reality on those with ME/CFS. Prejudice may determine what people perceive and in turn how they act towards people suffering from ME/CFS. As one person from our expert session said: "If one could only eliminate the burden of the shame that many people with ME/CFS live with and shift the focus to acceptance and to living their lives as best as possible, the quality of their lives would be so much better." The red edges drawn in Fig. 1 between the Understanding of ME/CFS and other nodes are also relevant for Prejudice as an actor in the network. However, the two are co-dependent: as the knowledge about, and understanding of, the ME/CFS increase, prejudice decreases.

The various actors have different roles and tasks, which are not predefined but evolved and given on the specific network of actors. In this network, the policies of governmental social services might result in rejection when the patient applies for governmental work assessment allowance, but with a patient with a different disease, this might be granted. This way the policies of governmental social services (which are the same for both patients) have a different impact on the lives of patients based on their diagnosis. The governmental social services caseworker might also have different predefined perceptions of ME/CFS compared to other diseases such as cancer. Another caseworker might have an entirely different perception of ME/CFS patients, which influence their judgment of the patient's need for governmental work assessment allowance.

The role and agency of the parent with a child diagnosed with ME/CFS changes based on the magnitude of their disease symptoms. For some, the disease debilitates the child from performing everyday tasks or staying in touch with friends. The parent could then take responsibility for some of these tasks, altering the parent-child relationship and their agency as a parent. The focus can shift from giving the child independence to a focus on the child as completely dependent on care from the parent. The delegation of tasks could cause the parent to go from not interfering in their child's social life to actively mediating their social interactions with friends by acting on behalf of their child.

Currently, the ME/CFS patients have to comply with the diagnostic criteria that apply to be diagnosed with ME/CFS. They also have to comply with policies of the State Educational Loan Fund and Governmental Social Services, to receive student loans or social benefits. These policies could be seen as a program of action, which the ME/CFS patients have to adhere to. These programs of actions are based on strong inscriptions. Therefore, it is not easy for ME/CFS patients to oppose these policies. This creates a big conflict in the network since the policies demand that one cannot study (not even part-time) and receives financial support from governmental social services. This is problematic for ME/CFS patients based on how their disease symptoms vary on a day-to-day basis. Young people with ME/CFS could often study if it were possible to do it at their pace. This should be provided without having severe economic consequences for the ME/CFS sufferer, and they should not have to deal with losing their student rights, such as they would today. Now people with ME/CFS encounter many roadblocks when trying to finalize their education. Therefore, there is a lack of alignment between the interests of those with ME/CFS and Governmental Social Services/ The State Educational Loan Fund. This lack is also visible in the current dire need for more medical research, treatment options and education of the general public to improve the general knowledge about the ME/CFS disease. Alignment means that the actors more or less share the same goals and motives.

To achieve stability, there has to be a process of compromising the interests of the welfare system and the ME/CFS patients. To make the network stable and aligned the result from the translation need to be agreed upon by all the actors. To increase knowledge about the disease might be an important step to remove some of the prejudice and these brick walls ME/CFS patients encounter in our society and to achieve alignments of the actors in the network. Further alignment could be archived by changing the policies of Governmental Social Services and providing flexibility and tailored solutions that better comply with the nature of the ME/CFS disease.

V. ANT AND GIGA-MAPPING

In the previous section, the research case is addressed by focusing on developing actor networks. The network describes divisive issues, the political and social dimensions. This is described by Storni (2015) as a part of the first turn in using ANT in design processes. The second part of this first turn is to look for ways to make these designed actor networks visible. Giga-mapping is an often used approach to making things visible in design processes, drawing upon design skills like making visual descriptions that foster communication and visual thinking. Giga-mapping is also a designerly way of dealing with super-complexity, which is highly relevant to our research case and its wicked problems. We argue that ANT and Giga-mapping complimented each other's strengths and weaknesses and used combined; it provides a powerful tool in sense-making processes. All relations among actors are significant for the initial sense-making process. This fits well with the Giga-mapping approach since it requires that the map includes everything, without ordering or giving personal meanings to the different elements involved. All the actors, explicitly or implicitly, contribute to the discussion around technology that could aid adolescents suffering from ME/CFS. ANT, on the other hand, was used as a lens to make connections and to group the different actors involved in important design spaces. Jointly, ANT and Giga-mapping provide a tool for critical reflection in the early sense-making phase of trying to understand adolescents with ME, and the situated design space related to technology that could be useful in their lives. In this research we, therefore, started by designing actor networks, then we used Giga-mapping to make these networks visible. The numerous challenges related to ME/CFS are not easy to solve and clearly relate to the main characteristics of wicked problems. We have proposed a holistic approach to sense-making when considering technology design for adolescents with ME/CFS. Using a literature review, a sense-making session with experts, ANT, and Giga-mapping, we gained insight into what ME/CFS is, developed the empathy with youth suffering from it. ANT provided rich descriptions of the complex design context and Giga-mapping visual representations of relations that emerged as relevant. They also helped to define possible areas of interest for design (Personal/family, Educational and Governmental). The outcome shows that our analysis aligns with a few studies from the literature that we have described: education and social arena are important. We initially found the Governmental space not as readily available to us as researchers/designers. However, one possible early design focus could be on increasing the general populations' knowledge about illness, thus helping to reduce prejudice and stigmatization. Based on our initial sense-making process, we argued for a focus on the Educational design space, both based on the current and future importance for the youth with ME/CFS and based on how the other children in their class can learn about ME/CFS. As mentioned, knowledge about ME/CFS is of utmost importance to reduce prejudices.

The developed Giga-map served three proposes for our research process. These relate to the questions of what to design? Who should be involved and aware of important issues? How should knowledge be developed and documented? Firstly, it helped us to increase our understanding of the design case and what to design for these users. Secondly, we also wanted to use the developed Giga-map as a design artifact that supports participation in the design of new technologies for youth suffering from ME/CFS. As mentioned, there is also limited previous research on technology design for ME/CFS sufferers; technology interventions have to be aware of and sensitive towards the manifestation of the ME/CFS illness and the attached political and social issues. Again, addressing the wickedness of the problem, the many challenges related to ME/CFS are not easy to fix and cannot be solved by any one single technological 'solution.' Instead, we focus on generating awareness within the community and the field of design by making things public through the Giga-map aiding to inform policy-makers and designers concerning youth suffering from ME/CFS. Using design as a powerful tool to make concerns about co-habitation and democracy visible. Fostering participation from relevant communities could prove to be the best way to address the wickedness of the problem. Thirdly, we want to use the initially developed map together with new Giga-maps and thick descriptions from the ANT analysis and our qualitative research, as a way of documenting design processes and technology developments addressing wicked problems.

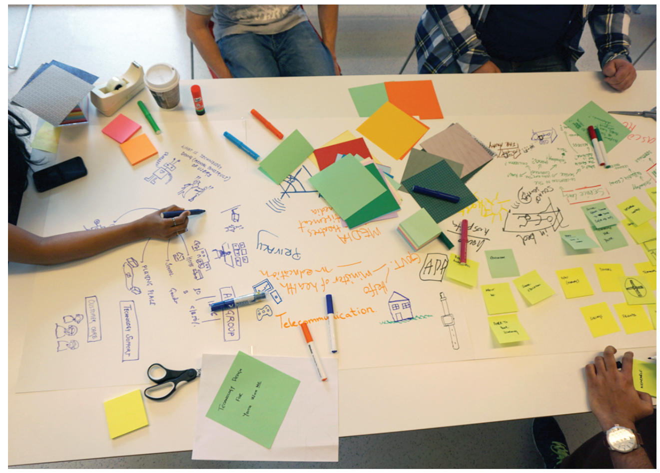


Figure 3. The Giga-map workshop.

VI. The Giga-Mapping Workshop

For the developed Giga-map to function as a means to document the pre-technology introduction design context, and in such a way that it could help other designers or researchers gain the knowledge we acquired through our sense-making process. It is imperative that the map is to be understood by others that do not necessarily possess prior knowledge about the design context. In fact, it is a question whether the Giga-map can be fully understood by other people than the ones who created them. To gain more perspectives on the design context and how others perceive our initial mapping of the design context we conducted a workshop, with four participants in addition to the researchers. It was immediately found that our map did not transfer the information to others that we had incorporated into the map. While as stated the map was useful for us in our initial sense-making, when we were trying to grasp and gain an understanding of the wicked problems in our research and design context, they were not successful as carriers of that knowledge to others. Keeping in mind that the Giga-map technique is not intended to 'tame' problems or be arranged or structured in a way that simplifies or in other ways bias the interpretation of reality. The rest of the workshop was used to explore how to make improvements.

We found that instead of showing the actors on our map by text and circles, it is much better to use icons or illustrations when possible. It was also stated that when trying to identify technology design opportunities the service lens from Design thinking could prove to be fruitful. In our next iteration of the Giga-map, we actively used these two findings.

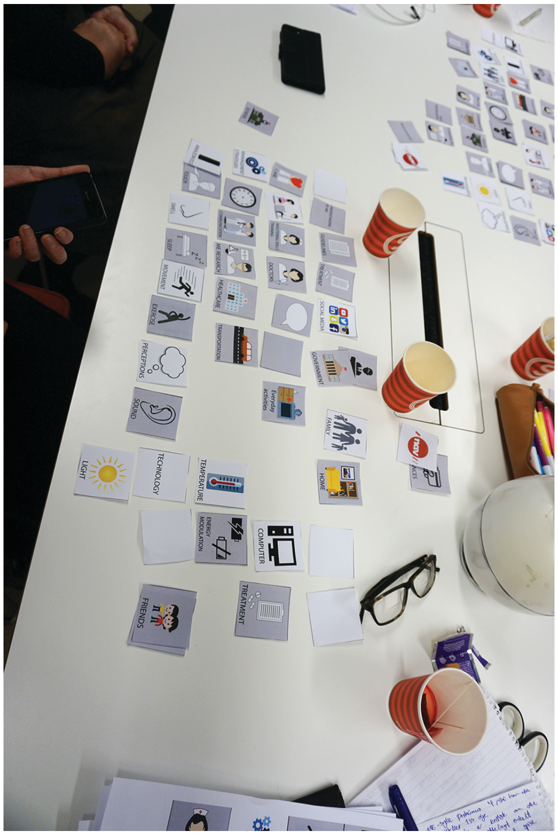


Figure 4. The robot from our first technology design case.

VII. The Second Expert Session

To validate the identified concepts and topics that should be included in the Giga-map, we performed a second expert session. The participants were the same as those who participated in the first session, except that the ME youth organization had a different representative participate in this second session. In this session, all the issues and concepts identified in the first expert session were made into cards inspired by [53][54][55]. The participants then sorted the cards and discussed the topics displayed on the cards. Even though we did validate that the most critical issues were included, we found the number of cards, the combined visual information, and the card sorting techniques to not fit well with the user group and the cognitive illness symptoms, mainly problems with memory and thought. To continue with the session, we had to start showing only one card at a time that was then discussed. This way we validated the importance of the identified topics. To get the participants to make an overall structure of all the cards that made sense regarding placement on the Giga-map was on the other hand not possible. We did, however, get much more information about the different topics and how they were relevant in their everyday living and life with the ME/CFS illness in general. The cards that especially generated much data in the interview was Governmental Social services, Private finances, Politics, Stigmatization, friends, ME support community, School, transportation, Education, Illness perception and knowledge in society, Everyday life activities, illness symptoms, energy modulation. The data collected were photographs, observation notes regarding the card-sorting method, and detailed notes recording the participant's answers made by two of the facilitators.

VIII. The First Technology Design Case

Our first technology design case focused on youth with ME/CFS, builds on all the insights from the initial sense-making process. The case involves the use of a little robot avatar to physically represents a child at school, giving a sick housebound child a view of the classroom in a non-demanding way, through the remote use of the robot supported by an iPad or smartphone app. This design possibility is being explored through collaboration with a start-up company called 'No Isolation,' who developed a robot avatar in parallel with our research on ME/CFS. The robot avatar aims to combat the harmful social isolation, which many young people experience due to illnesses or physical injuries. The youth with ME/CFS is a separate case of the research on the usefulness of the robot and its impact on children's lives. The hope was that the robot would prove to be a tool enabling them to fulfill their education, reduce their exposure to social isolation from their peers and, thus, give them a better starting point for their adult lives.

Figure 5. The robot from our first technology design case.

For a more thorough description of this study and our findings see [16]. The study involves nine participants that have used their own robot avatar placed in their homes and schools as part of our research. All participants except one that had an illness recovery are at the time of writing this article still using the robot avatar in its final commercial version. One finding from the robot avatar study is how beneficial the flexibility of instant access to the classroom is for youth with ME/CFS. A user having a day or even just an hour with decreased illness symptoms could spontaneously login and get some teaching and reduce their exposure to social isolation that day. This result relates to the finding in the first expert session of how education has to provide flexibility and that it is an advantage if the student with ME/CFS do not have to use energy on transportation to get access to classroom teaching. The flexibility of instant classroom access, illness accommodations to better fit with the energy modulation and the social benefits also is connected to the Educational design space and Personal/Family design space as shown in Fig. 2. However, this benefit also comes with a challenge related to the energy modulation, with a perfect balance between use and rest the technology could work in sync with the challenging nature of the energy modulation attached to the ME/CFS illness. There is, however, issues related to potential overuse and negative impact on illness symptoms. This mentioned possibility of a worsening of illness symptoms is one possible unintended harmful consequence of the technology introduction. Another unintended consequence could be if the users were pressured by others to overuse or if resources used to provide special education was withdrawn, a resource that was reported as successful for the youth with ME/CFS that has home tuition tailored to their educational needs. If such support were removed as feared by some parents in our study, the technology introduction would result in providing the users with a poorer educational offer. This is an example of how the technology introduction also affects the Governmental design space. Another example of how the Governmental design space is relevant for the robot avatar is that the Norwegian law was not formulated with the robot in mind. Thus, the placement and use of the robot in the classroom received many legal concerns, problems, and questions. Therefore, In December 2016 the Ministry of Education started working on addressing some of the issues surrounding the use of the robot avatars. They examined what the law says about the use and how to relate to it [56][57]. Furthermore, this effect also illustrates the robot avatars material agency, through the relations to other actors, the robot avatar has a capacity for action and influence on the network.

That politicians are supporting access to the technology for everyone, is crucial in providing the technology to families with low-incomes or parents that have reduced incomes since they are providing care for their loved ones suffering from ME/CFS. If the cost of the robot and its use is too high, it could result in increased class differences between families of low and high income. We observe that some parents and physicians of the users in our study, have taken the initiative and contacted the media and politicians to support access to the technology for others. A new actor in the network is the commercial startup company developing the robot. Being part of a business, they have to maintain high enough unit price for the robots not to go bankrupt. Therefore, there is a negotiation between the needs of the families of ME/CFS sufferers, which need the price low enough so that the technology is accessible to everyone that needs it and the company that needs profit to stay viable. This negotiation of the cost also involves the possibility of politicians making the robot available through the welfare system in Norway. As mentioned there are current efforts done to convince the politicians to take action and include the robot in our welfare system. These efforts come from people within ME organizations, a user's pediatric physician, teachers, and parents.

Regarding the stigmatization of people with ME/CFS, we have not gotten any reports from our users stating that these robots were perceived as stigmatizing to use or own, more specifically they were reported by peers as cool or neutral. For people within the ME/CFS community, it was perceived as particularly positive that youth suffering from ME/CFS was included in the study, in the same way as children suffering from less stigmatized illnesses like cancer. Seeing how popular the company has become for young people, how the study puts focus on the importance of providing technology support for young people with ME/CFS, could help increase the understanding of ME/CFS in our society.

To sum up, for the technology introduction to be successful for the users, it is critical that other actors are enrolled like teachers, parents, and politicians. Keeping the energy modulation in mind and together with the users, they all have to contribute to providing a successful technology introduction outcome. Our case illustrates how a technology seen, as a new actor in the network, becomes a catalyst of change affecting many of the actors in the pre-established network, changing roles and introduce new actors. The technology aims to mainly focus on providing users with access to education and reduce social isolation from their peers. A deeper more thorough ANT inspired analysis uncovers how the introduction of this technology sparks and in some cases demands changes, affecting many more of the actors in the network than intended by the technology development initially targeting the Educational design space. In fact, we saw that the robot avatar has a strong presence in all the design spaces displayed in our Giga-map.

The process of developing the Giga-map supported by ANT resulted in a productive process where our understanding of the design context was simultaneously enriched. During our first technology introduction in this design contest, the developed map provided the recognition of invisible issues and possible problems that should be addressed in the redesign of the robot or the adopted strategies for use. The awareness of matters such as a low understanding of the illness or possible issues related to stigmatization. Served as a constant reminder to ask difficult questions during research and when envisioning further use. Both concerns are in nature human and all humans involved could be possible carriers of attitudes or wrongful understanding of ME/CFS that could result in robot users being stigmatized or to get worse if the energy modulation is not attended to. Essential questions then became what power does the various human actors in the design context have about the user or the use of the robot? How could the robot or the interaction be designed to lower the chance of such ill effects?

IX. Conclusion

We have argued that sense-making is a crucial step in understanding complex design domains in HCI, as exemplified by the task of identifying technology design opportunities aimed to support everyday life of adolescents with ME/CFS. Published research results and ethnographic methods were used to understand risks associated with having ME/CFS. ANT and Giga-mapping were used to analyze and visualize input from sessions with experts. The ANT analysis primarily made us aware of the problems related to the low level of knowledge about ME/CFS in our society, and how those having the illness risk stigmatization. It became clear that numerous problems adolescents with ME/CFS face in our society cannot be solved by one technological 'solution,' but instead have to be addressed in a holistic way, by several actors including families, schools, peers, politicians, health-care providers, and research communities and others. The use of Giga-mapping helped to visualize all these crucial actors, including the authors, and relations among them. The utilization of rich descriptions and Giga-maps throughout the design process is also a way of annotating, documenting, reflecting on and sharing of divisive concerns and lessons learned. As this is a messy and unstructured process, we need to mention that the communicative value of Giga-maps that we produce to others than ourselves, is very low. The map had to be visually re-interpreted to make essential issues from the sense-making analysis approach visible (understandable) for the HCI research community. We, therefore, encourage some testing of the visual representation to make sure that it conveys the information as intended.

Our analysis uncovered three main design spaces relevant to our research case: educational, government and Personal/Family. While there are many different relations between actors within these spaces, and different levels of granularity that can be represented, when it comes to technological solutions these seem to be rather well separated, indicating many opportunities and challenges when working on boundaries between them. Further, we argue that these design spaces can aid in framing how potential new research projects, design concepts or technology artifacts relate to issues within the situated design space.

We believe that the approach, as presented in this paper, successfully combined methods and techniques that can guide HCI designers and researchers to conduct a broad inquiry and sense-making for new, complex design domains, where it is not easy to find solutions that work for all. Instead, the approach offers a way to gain a profound and vital understanding of the problem space.

ACKNOWLEDGMENT

Enormous gratitude goes to the Norwegian ME organization, the ME youth organization, No Isolation and all participants in expert sessions and the workshop for a wonderful collaboration in this project.

REFERENCES

[1] J. Børsting and A. L. Culén, "Sense-Making in Complex Healthcare Domains: The Role of Technology in Every Day Lives of Youth Suffering from ME/CFS," Presented at the eTELEMED 2016, The Eighth International Conference on eHealth, Telemedicine, and Social Medicine, 2016, pp. 236–242.

[2] B. W. Head and J. Alford, "Wicked Problems Implications for Public Policy and Management," Adm. Soc., vol. 47, no. 6, pp. 711–739, Aug. 2015. 1955.

[3] S. Waddock, "The Wicked Problems of Global Sustainability Need Wicked (Good) Leaders and Wicked (Good) Collaborative Solutions," J. Manag. Glob. Sustain., vol. 1, no. 1, pp. 91–111, Jun. 2013.

[4] R. Buchanan, "Wicked Problems in Design Thinking," Des. Issues, vol. 8, no. 2, pp. 5–21, Apr. 1992.

[5] H. W. J. Rittel and M. M. Webber, "Dilemmas in a general theory of planning," Policy Sci., vol. 4, no. 2, pp. 155–169, Jun. 1973.

[6] J. Kolko, Austin Center for Design, Wicked problems problems worth solving. Austin, Texas: Ac4d, 2012.

[7] J. Redström, "Towards user design? On the shift from object to user as the subject of design," Des. Stud., vol. 27, no. 2, pp. 123–139, Mar. 2006.

[8] J. M. Perrin, L. E. Anderson, and J. V. Cleave, "The Rise In Chronic Conditions Among Infants, Children, And Youth Can Be Met With Continued Health System Innovations," Health Aff. (Millwood), vol. 33, no. 12, pp. 2099–2105, Dec. 2014.

[9] E. A. Witt, A. J. Massman, and L. A. Jackson, "Trends in youth's videogame playing, overall computer use, and communication technology use: The impact of self-esteem and the Big Five personality factors," Comput. Hum. Behav., vol. 27, no. 2, pp. 763–769, Mar. 2011.

[10] R. L. Punamäki, M. Wallenius, C. H. Nygård, L. Saarni, and A. Rimpelä, "Use of information and communication technology (ICT) and perceived health in adolescence: The role of sleeping habits and waking-time tiredness," J. Adolesc., vol. 30, no. 4, pp. 569–585, Aug. 2007.

[11] A. L. Culén and A. Karpova, "Designing with Vulnerable Children: a Researchers Perspective," in Human-Computer Interfaces and Interactivity: Emergent Research and Applications Book, IGI Global, 2014, pp. 118–136.

[12] A. L. Culén and M. van der Velden, "The Digital Life of Vulnerable Users: Designing with Children, Patients, and Elderly," in Nordic Contributions in IS Research, Springer Berlin Heidelberg, 2013, pp. 53–71.

[13] B. Latour, "On using ANT for studying information systems: a (somewhat) Socratic dialogue," Soc. Study Inf. Commun. Technol., pp. 62–76, 2004.

[14] C. Storni, "Notes on ANT for designers: ontological, methodological and epistemological turn in collaborative design," CoDesign, vol. 11, no. 3–4, pp. 166–178, Oct. 2015.

[15] P. K. Chilana, A. J. Ko and J. Wobbrock, "From User-Centered to Adoption-Centered Design: A Case Study of an HCI Research Innovation Becoming a Product," in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, New York, NY, USA, 2015, pp. 1749–1758.

[16] J. Børsting and A. L. Culén, "Experiences with a Research Product: A Robot Avatar for Chronically Ill Adolescents" Smart Technology Applications in Business Environments. IGI Global. 8. pp. 159 - 183, 2017.

[17] A. L. Culén and A. Følstad, "Innovation in HCI: What Can We Learn from Design Thinking?," In Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational, NY, USA, 2014, pp. 849–852.

[18] A. Gasparini, "Perspective and Use of Empathy in Design Thinking," in ACHI 2015, 2015.

[19] N. Afari and D. Buchwald, "Chronic Fatigue Syndrome: A Review," Am. J. Psychiatry, vol. 160, no. 2, pp. 221–236, Feb. 2003.

[20] G. J. Bringsli, A. Gilje, and B. K. G. Wold, "The Norwegian ME Association National Survey Abridged English Version." The Norwegian ME Association, 2014.

[21] Ø. Fluge et al., "Benefit from B-Lymphocyte Depletion Using the Anti-CD20 Antibody Rituximab in Chronic Fatigue Syndrome. A Double-Blind and Placebo-Controlled Study," PLoS ONE, vol. 6, no. 10, e26358, Oct. 2011.

[22] V. R. Anderson, L. A. Jason, L. E. Hlavaty, N. Porter, and J. Cudia, "A review and meta-synthesis of qualitative studies on Myalgic Encephalomyelitis/chronic fatigue syndrome," Patient Educ. Couns., vol. 86, no. 2, pp. 147–155, Feb. 2012.

[23] S. M. van Geelen, G. Sinnema, H. J. M. Hermans, and W. Kuis, "Personality and chronic fatigue syndrome: Methodological and conceptual issues," Clin. Psychol. Rev., vol. 27, no. 8, pp. 885–903, Dec. 2007.

[24] A. Winger, M. Ekstedt, V. B. Wyller, and S. Helseth, "'Sometimes it feels as if the world goes on without me': adolescents' experiences of living with chronic fatigue syndrome," J. Clin. Nurs., vol. 23, no. 17–18, pp. 2649–2657, Sep. 2014.

[25] P. White et al., "Comparison of adaptive pacing therapy, cognitive behaviour therapy, graded exercise therapy, and specialist medical care for chronic fatigue syndrome (PACE): a randomised trial," The Lancet, vol. 377, no. 9768, pp. 823–836, Mar. 2011.

[26] Norwegian ME Association (in Norwegian only): me-foreningen.com/. Accessed on April 12th, 2016.

[27] H. Holman and K. Lorig, "Patient self-management: a key to effectiveness and efficiency in care of chronic disease," Public Health Rep., vol. 119, no. 3, pp. 239–243, 2004.

[28] J. M. Corbin and A. Strauss, "Unending work and care: Managing chronic illness at home," vol. xviii. San Francisco, CA, US: Jossey-Bass, 1988.

[29] K. R. Lorig and H. Holman, "Self-management education: history, definition, outcomes, and mechanisms," Ann. Behav. Med. Publ. Soc. Behav. Med., vol. 26, no. 1, pp. 1–7, Aug. 2003.

[30] E. G. Dowsett and J. Colby, "Long-Term Sickness Absence Due to ME/CFS in UK Schools," J. Chronic Fatigue Syndr., vol. 3, no. 2, pp. 29–42, Jan. 1997.

[31] M. E. B. Smith et al., "Treatment of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: A Systematic Review for a National Institutes of Health Pathways to Prevention Workshop," Ann. Intern. Med., vol. 162, no. 12, pp. 841–850, Jun. 2015.

[32] A. Farmer, T. Fowler, J. Scourfield, and A. Thapar, "Prevalence of chronic disabling fatigue in children and adolescents," Br. J. Psychiatry, vol. 184, no. 6, pp. 477–481, Jun. 2004.

[33] National Collaborating Centre for Primary Care (UK), "Chronic Fatigue Syndrome/Myalgic Encephalomyelitis (or Encephalopathy): Diagnosis and Management of Chronic Fatigue Syndrome/Myalgic Encephalomyelitis (or Encephalopathy) in Adults and Children," London: Royal College of General Practitioners (UK), 2007.

[34] B. M. Carruthers et al., "Myalgic Encephalomyelitis: International Consensus Criteria," J. Intern. Med., vol. 270, no. 4, pp. 327–338, Oct. 2011.

[35] F. J. Grajales III, S. Sheps, K. Ho, H. Novak-Lauscher, and G. Eysenbach, "Social media: a review and tutorial of applications in medicine and health care," Journal of medical Internet research, vol. 16, no. 2, 2014 Available: http://www.jmir.org/2014/2/e13/. [Accessed: 18-Jan-2016].

[36] K. Best and S. Butler, "Second life avatars as extensions of social and physical bodies in people with Myalgic Encephalomyelitis/Chronic Fatigue Syndrome," Continuum, vol. 27, no. 6, pp. 837–849, Dec. 2013.

[37] K. Best and S. Butler, "The Necessity of Control for Computer and Internet Users with Myalgic Encephalomyelitis/Chronic Fatigue Syndrome," Inf. Commun. Soc., vol. 16, no. 7, pp. 1042–1060, Sep. 2013.

[38] A. K. Sheridan and members of the C. P. G. E. of C. with M. (Scotland), "Exploring E-learning provision for Children with ME in Scotland," Educ., vol. 2, no. 1, pp. 78–80, Jan. 2013.

[39] T. Brown, "Design Thinking," Harv. Bus. Rev., vol. 86, no. 6, pp. 84, 2008.

[40] T. Brown and J. Wyatt, "Design Thinking for Social Innovation (SSIR)," Stanford Social Innovation Review, 2010.

[41] U. Johansson-Sköldberg, J. Woodilla, and M. Çetinkaya, "Design Thinking: Past, Present and Possible Futures," Creat. Innov. Manag., vol. 22, no. 2, pp. 121–146, 2013.

[42] B. Sevaldson, "Giga-Mapping: Visualisation for complexity and systems thinking in design," Nordes, vol. 4, Mar. 2011.

[43] O. Hanseth, M. Aanestad, and M. Berg, "Guest editors' introduction: Actor-network theory and information systems. What's so special?," Inf. Technol. People, vol. 17, no. 2, pp. 116–123, 2004.

[44] B. Latour, "Aramis, or, The love of technology," Cambridge, Mass: Harvard University Press, 1996.

[45] J. Law, "Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity," Syst. Pr., vol. 5, no. 4, pp. 379–393, Aug. 1992.

[46] A. Tatnall and A. Gilding, "Actor-Network Theory in Information Systems Research," 2005.

[47] Y. Akama, "Being awake to Ma: designing in between-ness as a way of becoming with," CoDesign, vol. 11, no. 3–4, pp. 262–274, Oct. 2015.

[48] A. Yaneva, "Making the Social Hold: Towards an Actor-Network Theory of Design," Des. Cult., vol. 1, no. 3, pp. 273–288, Nov. 2009.

[49] D. Stuedahl and O. Smørdal, "Matters of becoming, experimental zones for making museums public with social media," CoDesign, vol. 11, no. 3–4, pp. 193–207, Oct. 2015.

[50] J. Schoffelen et al, "Visualising things. Perspectives on how to make things public through visualisation," CoDesign, vol. 11, no. 3–4, pp. 179–192, Oct. 2015.

[51] M. Aanestad, "The camera as an actor design-in-use of telemedicine infrastructure in surgery," *Computer Supported Cooperative Work (CSCW)*, vol. 12, no. 1, 2003, pp.1-20.

[52] E. Sayes, "Actor–Network Theory and methodology: Just what does it mean to say that nonhumans have agency?," Social Studies of Science, vol. 44, no. 1, pp. 134-147, Feb. 2014.

[53] W. Hudson, "Playing your cards right: getting the most from card sorting for navigation design," interactions, vol. 12, no. 5, pp. 56-58, 2005.

[54] A. L Culén and M. van der Velden, "Making Context Specific Card Sets-A Visual Methodology Approach: Capturing User Experiences with Urban Public Transportation," International Journal on Advances in Intelligent Systems, *8*(1&2), pp. 17-26. 2015.

[55] M. M. Sommervold and M. van der Velden, "Transition cards: Designing a method with and for young patients," IADIS International Journal on Computer Science & Information Systems, vol. 10 no. 2, 2015.

[56] P4 (in Norwegian only): www.p4.no/nyheter/--problematisk-med-robot-i-klasserom/artikkel/689010/. Accessed on May 29th, 2018.

[57] The Norwegian Ministry of Education (in Norwegian only): https://www.udir.no/regelverk-og-tilsyn/finn-regelverk/etter-tema/Skoleeiers-ansvar/bruk-av-robot-i-undervisningen/. Accessed on May 29th, 2018.

Architectural Backbone Evaluation for Data Stream Processing within the WINNER DataLab A Project Focused Point of View

Sebastian Apel*, Florian Hertrampf†, and Steffen Späthe‡

Chair for Software Engineering
Friedrich-Schiller-University Jena
D-07745 Jena, Germany
Email: sebastian.apel@uni-jena.de*, florian.hertrampf@uni-jena.de†, steffen.spaethe@uni-jena.de‡

***Abstract*—Smart platforms for the integration of sensor and actor networks require collecting, analysing and evaluating data. Our research project WINNER aims to integrate systems which cover electromobility, energy consumption within residential areas, local energy production, e. g., with photovoltaic systems, and storages for local smart grids. WINNER wants to use such a platform to optimise the energy consumption. Every actor within a residential area has to be considered, and integration into a centralised data stream process is necessary. As a non-hard real-time system, the platform has to solve enterprise application integration problems, looking at complex event processing and knowledge discovery in data. This paper targets to analyse possible architectural backbone technologies. Out of a wide range of potential technologies Node-RED, Apache NiFi and Apache Camel are selected and compared. Those technologies with a diverse field of application are used to implement a comparable test setup. Furthermore, they are analysed through their characteristics of processing, execution, usability and simplicity. As measured, Node-RED, Apache Camel and Apache NiFi indicate stable and fast message processing, especially in the case of raising message throughput. Node-RED surprises with constant memory and CPU loads and seems to be an exciting option in rapid prototyping.**

***Keywords*–System Architecture; Stream Processing; Message Routing; Complex Event Processing; Renewable Energy; Smart Grid.**

I. INTRODUCTION

Local energy networks in modern residential areas are used by additional installations such as charging stations for electric vehicles, local energy production (e.g. with photovoltaic systems) and energy storage systems. The installations could operate independently and contribute. However, an increase in efficiency would be conceivable if those installations worked in a coordinated manner. An integration platform required for this could collect and process information from all installations and influence an entire system to operate an optimised overall setup. Our research project "Wohnungswirtschaftlich integrierte netzneutrale Elektromobilitat in Quartier und Region" (WINNER) [2] aims to integrate such systems and wants to use such a platform to optimise the local energy consumption.

Therefore, three main tasks have to be considered and brought together by our so-called WINNER DataLab (WDL). At first, we collect all the produced data and store them in a meaningful way. After that, potentials have to be found, e. g., correlating weather forecasts, electricity consumption, specific time information, and the usage characteristic of electric vehicles (EVs) to optimise external energy purchase for charging batteries. In the end, we have to optimise operation. So we could control and accumulate electric energy locally or supply it to the grid. Maybe it is superior or necessary to get energy from another grid operator, e. g., in case of too less output of the local energy production.

The facts as mentioned earlier imply an information flow managed by data streams. These must be routed and checked for mistakes. Beyond various data sources have to be integrated, like Representational State Transfer (REST) interfaces based on Hypertext Transfer Protocol (HTTP) or mail services, as well as other proprietary transport and communication protocols.

According to backbone technologies, we have to discuss the potentials of tools made for message routing and analysing within the WDL. We focus on event-based approaches and easy integration of external components. In the end, we ask for a tool that offers the possibility of routing messages and analysis of the contained data without dropping information. For the decision-making process, a unified prototype was realised in a selection of possible tools, compared and considered under load. This publication summarises our decision process. We want to explicitly outline, that our aim is not to benchmark the preselected tools in detail. We want to observe the memory usage development as well as the CPU consumption of the tools while processing an increasing amount of messages. So, we can get an idea of the behaviour of the message processing systems.

In Section II, related work about terms and projects related to our approach are discussed. Section III presents the level 0 view of our WDL, and the following Section IV lists the requirements we impose on this system. As a result of that, a short overview of possible tools is presented in Section V. Furthermore, three tools are used to implement and compare a uniform task in Section VI and analysed them in Section VII by using measurement values of latency, memory consumption and CPU load. Finally, we discuss the results in Section VIII.

II. TERMS AND RELATED WORK

The WDL seems to be far away from traditional database management systems. The WDL should be usable as a platform for data scientists for mining knowledge as well as a platform to attach analyses for already known behaviours and relationships directly on data streams. Furthermore, the WDL should consume data from different kinds of systems as well as produce data to optimise the usage of those systems.

Connecting various types of applications by using their provided data and processes belongs to the term of enterprise application integration (EAI) [3, P. 3]. Within the area of application integration terms like message-oriented middleware (MOM) and service-oriented architecture (SOA), as well as enterprise service bus (ESB), describe how to challenge those use cases [4, S. 1][5]. As a traditional approach, MOM describes how to use asynchronous messages to decouple applications based on messaging systems [5]. SOA, on the other hand, represents an architectural concept where applications publish their precisely defined functionalities within reusable services [5]. Finally, ESB draws an open standard, which merges those ideas and defines a distributed architecture used to integrate applications. The architecture itself describes calls and distribution of messages between integrated applications [5].

Within the scope of EAI, the enterprise integration patterns (EIPs) are the base of tools to solve integration problems. The EIPs describe a set of reusable patterns without a particular technology reference. Base concepts within these patterns are the usage of “routing” and “messages” [6].

Beside the application integration itself, activity tracking, sensor networks and analysing of market data is a central topic within the so-called complex event processing (CEP). CEP describes a general term for methods, techniques and tools. CEP helps to process events while they happen [7, S. 163].

Bringing together EAI, MOM, SOA, ESB and CEP seem to be not distinctly possible. Currently, there are multiple terms to describe the problem of integration, routing, processing and analysing. The first one, Information Flow Processing, is described in [8]. This term focuses on event processing in combination with data management to “collect information produced by multiple, distributed sources, to process it in a timely way” [8]. Another term, streaming data system, focuses on processing data streams within “a non-hard real-time system that makes its data available at the moment a client application needs it” [9].

While EAI, MOM, SOA, ESB and CEP are concepts to assemble setups based on already known behaviours and relationships between messages (or events), data sciences utilise tools to mine knowledge based on already available data. This part within the WDL uses concepts from knowledge discovery in databases (KDD), which describes methods to statical analyses, applications within the field of artificial intelligence (AI), pattern recognition and machine learning [10, S. 3].

In addition to this classification of concepts within our field of application, related work targeting onto architectural drafts in data grids and smart grids can be used. Chervenak et al. [11], e. g., describes basic principles for designing data management architectures and Tierney et al. [12] introduce concepts how to monitor such grids. Furthermore, Appelrath et al. describe in [13] the process of developing an IT-architecture for smart grids as a result of a German research project, and Rusitschka et al. [14] present a computing model for managing real-time data streams of smart grids within the scope of the energy market. But, these approaches are not directly applicable to our use case. Either they are large-scale, or focusing on data storing and mining. However, they can be considered within our architecture, which has to fill the gap between smart grids, data storing, possibilities for data mining as well as non-hard real-time event processing.

Besides the mentioned, more general architectures for the internet of things (IoT) are discussed in [15] and [16]. They point out that a flexible layered architecture is necessary to connect many devices. As an example, classical three- and five-layer architectures are mentioned here, e. g., as used within [17] and [18], as well as middleware or SOA-based architectures. Besides this discussion about architectures, [19] demonstrates the integration of IoT devices within a service platform which uses the micro-service architecture for this approach, which can be understood as a specific approach for SOA [20, P. 9].

In addition to IoT, measurement systems for IoT can also be considered. They also have to integrate different end systems. Further, they have to record different measured values and provide interfaces for analyses and calculations. For this, approaches like SOA or event-driven architecture (EDA) can be taken up, as demonstrated in [21]. This approach uses SOA and EDA in combination with an ESB. Using the micro-service architecture can be seen in [22] and [23] as loosely coupled by using the EAI [3, P. 3]. They describe a reference architecture using micro-services for measurement systems, which connects required data adapters as well as calculation and storage services, one more time through an ESB.

III. Architectural Draft

The WDL is a platform to gather, analyse and provide information to optimise the operation of the WINNER-project setup. Until now, this platform seems to be a data streaming platform which has to solve various integration tasks. At this point, the level 0 view of the WDL is discussed. Taking a look at the data sources and data sinks help to get a better understanding of what the WDL should do.

At first, there are external services. They are sending messages to the WDL, or it acquires data from them. These data packages have to be assumed as heterogeneous, e. g., carsharing data of a booking system the WDL is connected to. That means the WDL gets information on bookings like start time and end time. Out of that, current state updates on a reservation such as an earlier beginning or a defect vehicle can be received. Another data source offers messages containing information on the current electrical power consumption. The actual electricity price is obtained by an interface of European Energy Exchange (EEX). Data of photovoltaic systems or batteries are gained through System, Mess- und Anlagentechnik (German solar energy equipment supplier) (SMA) interfaces. The German Meteorological Service [24] offers historical information on the past weather, an application programming interface (API) of the online service OpenWeatherMap [25] is available for weather forecasts.

A system working with time series, forecasts and master data must be created. As visualised in Figure 1, the WDL is positioned between the sources above, and at least five data sinks. These refer to controllable devices like a charging station, a battery or Smart Home systems. On the other hand,

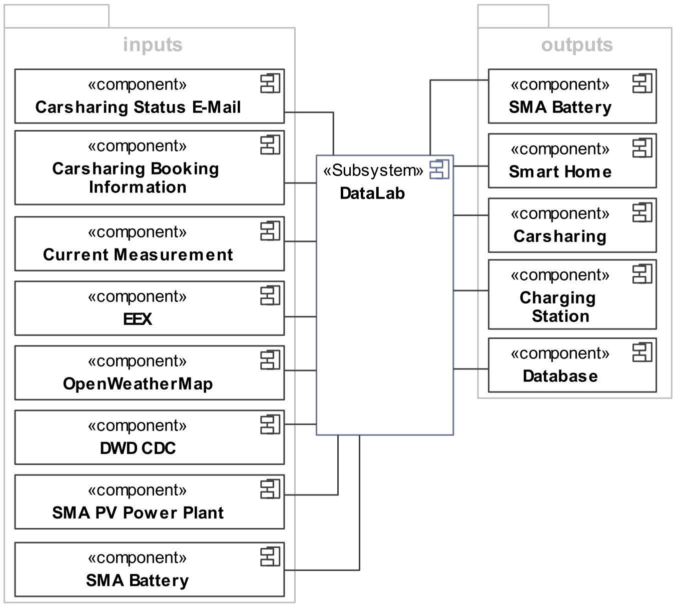


Figure 1. Level 0 view of the WDL.

data is delivered to the carsharing service and dumped to a database. Within our setup, a KairosDB is used as our data storage interface as it is suitable for working with larger time series and quite easy to use in combination with a Cassandra backend for storage.

An unanswered question is how the different components can be integrated and how analysis, as well as event processing, can be handled. The WINNER project focuses on the integration of components of the residential area into the Smart Grid. That means predictions must be made to get an overview of the future power consumption and the electricity production. Either one charges the batteries or one uses the stored energy to overcome load peaks. The prediction mechanism might be implemented by using artificial neural networks (ANNs) or regression methods. Thinking about energy production predictions, it might be neccessary to receive information from hardware components like SMA-devices and weather services. These specific data formats require reshaping to use them in prediction mechanism. Using input filters, output filters, and stream routing the arriving information is transformed and sent to the prediction and dump units. These units save the data, send commands or just forward data to external devices.

IV. Requirements

One can divide up the list of requirements into three subsets. The first one refers to the system in general; the second touch the various components and the third covers the aspects of architecture and functional groups.

Thinking of the system in general shows that the ability to *process time series data* is required. An incoming message contains a time value referring to a point and a value, e. g., the result of a measurement. The WDLs task on an incoming message is to associate the arriving values with a data source. Possible data sources are photovoltaic installations, batteries, power consumption measurement devices or actual weather data. Out of that, the system must handle forecast data. They are unique because a complete time series and a time value, which refers to a validity point are included. At the time this point describes the time series is valid. Contemplable data sources are weather forecast services or EEX. The last category covers master data without time dependencies. Booking information or general data on devices and services belong to this group. This data may be very unstructured like text-only entries.

Focusing on technical aspects derived from our architectural draft in Section III, the WDL needs the ability to *process JSON, XML and CSV values*. Out of that proprietary formats have to be handled as well. Especially photovoltaic and smart metering installations tend to send production data in proprietary formats.

Central non-functional requirements are *scalability and reliability*. The latter refers to interfaces receiving data from external services and devices. On occurring errors incoming and shortly arrived messages should not get lost. Referring to the message rates previous estimations showed, that our system has to process about 30 messages per second on average. Thinking of peak times, we estimate a message occurrence higher than the average by a factor of 10. Especially, if external information-providing services use caches to buffer data and send them shortly afterwards or in case of network problems this may happen.

Keeping the interfaces in mind, one has to think of the *necessary contact points to other services* or the environment in general. The consumer interfaces of the WDL have to accept HTTP requests, especially while communicating with REST services. Similarly, FTP servers must be communicated with. The WDL must receive and process e-mails as well. Likewise, a file-based data transfer is needed. Finally, there are interfaces to external services using proprietary communication formats via TCP or UDP. The developed system has to enable the reception of messages sent by them. In contrast, the message producing components of the WDL primarily need to communicate via HTTP. Particularly the interface to a database can be made up of simple REST client services sending HTTP-based messages.

After paying attention to input and output components, the *internal processes of routing and filtering* shall be characterised. Asynchronous processing describes an essential requirement. Message queues or small buffer databases may decouple various components so they can work without waiting for each other to terminate. Furthermore, incoming messages caused by occurring events have to be converted into an internal format. To achieve this, the WDL can extend these data packages with additional information. However, after processing unneeded contents must be removed as well. Alongside external descriptors have to be mapped to internal descriptors and vice versa.

The WDL has to *transform the incoming data into an internal format* for further processing. Additionally, the WDL has to be capable of providing data for doing manual statistical evaluations and analysis. Furthermore, the WDL has to be capable of triggering automatic evaluations and forecasts as additional components. This work is done while keeping the CEP pattern in mind.

Within this paper, we leave out the specific aspect of data storage. That means different databases are not discussed or compared. The built prototype uses a KairosDB to persist

TABLE I. TOOL OVERVIEW AND CLASSIFICATION. CLASSIFICATION IS BASED ON TOOLS TO HANDLE EAI, CEP KDD AND RE.

Name	EAI	CEP	KDD	RE
Apache Camel	✓	✗	✗	✗
Apache Storm	✗	✓	✗	✗
Apache Spark	✗	✗	✓	✗
Apache Hadoop	✗	✗	✓	✗
Apache ServiceMix	✓	✓	✗	✓
Apache NiFi	✓	✗	✗	✗
Siddhi	✗	✓	✗	✗
ESPER	✗	✓	✗	✗
WSO2 CEP	✗	✓	✗	(✓)
RapidMiner	✗	✗	✓	✗
KNIME	✗	✗	✓	✗
Node-RED	✓	✗	✗	✗
JBoss Fuse	✓	✓	✗	✓
StreamLine	✗	✓	✗	✗
Hortonworks SAM	✓	✓	✗	✓

time series data. It was chosen because of an existing simple HTTP-based interface that provides easy access.

Furthermore, security and data protection with regard to the sensitive information collected is not taken into account in this analysis. The efforts required for this influence the resulting data throughput but are not relevant for this evaluation.

V. TOOL OVERVIEW

The WDL requirement analysis illustrates an EAI task with KDD topics. Furthermore, results gathered from KDD could result in CEP related tasks, which have to be considered as well. The following list of tools covers these tasks. Of course, this list is not complete. There are a lot of tools available to handle specific tasks within the area of EAI, KDD or CEP. Our selection focuses on widely used, platform independent and easily accessible tools with suitable licenses models. Thus, the list of our selection contains mainly open source tools.

Selected tools will be classified into at least one of our primary topics: (1) tools to handle KDD related tasks, (2) tools to solve EAI related tasks and (3) tools to implement CEP related tasks. Additionally, there are (4) tools providing runtime environments for executing solutions solved with tools from class (1), (2) and (3).

Table I lists our selection of considered tools. Furthermore, this table classifies them within our previously identified main topics. Apache Camel is an open source lightweight framework to solve EAI problems based on an implementation of EIPs in [6]. Furthermore, a lot of components are available to extend the functionality of Apache Camel [26]. Apache Storm is an "open source distributed realtime computation system" with a lot of use cases like "realtime analytics, online machine learning, continuous computation". This scalable environment can handle a lot of data streams within a specific Storm topology [27]. Apache Spark [28] and Apache Hadoop [29] are tools for knowledge discovery in data. They differ in performance as well as their internal approaches in data storage and processing. Apache Service Mix [30] and JBoss Fuse [31] are integration containers, which include other tools like Apache Camel. Apache NiFi [32] provides a web interface for configuring directed graphs, which represent the processing of incoming messages. It uses so-called flow files for information representation. Siddhi [33] and ESPER [34] are CEP engines and can be used as standalone tools as well as the integration of tools like Apache Camel. WSO2 CEP is a runtime environment for the CEP engine Siddhi, which adds user interfaces for external and internal usage [35]. RapidMiner [36] and KNIME [37] are tools for knowledge discovery in already existing data. It is also possible to integrate interfaces to access data streams and use a wide range of algorithms to analyse collected data. Finally, Node-RED is a message processing framework with IoT roots and can be used to solve application integration problems quickly. This framework is based on Node.js, can be extended with additional packages and deployed into cloud services like Bluemix [38]. StreamLine and Hortonworks Streaming Analytics Manager (SAM) are aimed to develop and deploy Streaming Analytics applications visually. Therefore, Hortonworks SAM and StreamLine provide bindings for different streaming engines, a rich set of streaming operators as well as operational lifecycle management [39].

VI. PROTOTYPE

According to [1] we have selected three tools based on our preselection in Section V and our experience. We want to use the tools within a uniform test setup. This selection focuses on tools from different fields of application: (1) Node-RED because of its simplicity within the field of IoT, (2) Apache NiFi because of its easily configurable data flow and (3) Apache Camel as the reference implementation for EIPs in combination with Wildfly as Java EE based runtime environment.

The comparison is done with a uniform test setup with a simplified task. This setup combines the integration of a REST-based data source which encodes data with JSON, a KairosDB based data sink with HTTP interface which consumes JSON encoded data as well, and a calculation of the mean according to a particular sender device, e. g., a photovoltaic station, has to be calculated across multiple messages. The time window of these multiple messages is ten seconds. That means if sender "Station A" sends a message at 10:00:00 am the values "Station A" sent between 09:59:50 am and 10:00:00 am are used for calculating the mean. The result is delivered via HTTP request to an external service which consumes JSON encoded data as well as the data source and KairosDB.

The data source in our test setup gets its messages from a generative photovoltaic data endpoint in configurable intervals. This source device transmits structured data like the tuple "(time, energy, station, id)". The first value of the generated data tuple represents a long value as a point in time, the second value a double based energy value of solar insolation. Furthermore, a tag containing the string based station name is included. Finally, the last value is a string based identifier of this single message for further time measurements. The identification value does not contain any relevant information in the context of energy data aggregation. It is only used to register and match the outgoing and incoming messages on the peripheral systems around the measurement environment.

The KairosDB endpoint of this setup gets its message as structured data like the tuple "(name, value, tags, time, id)".

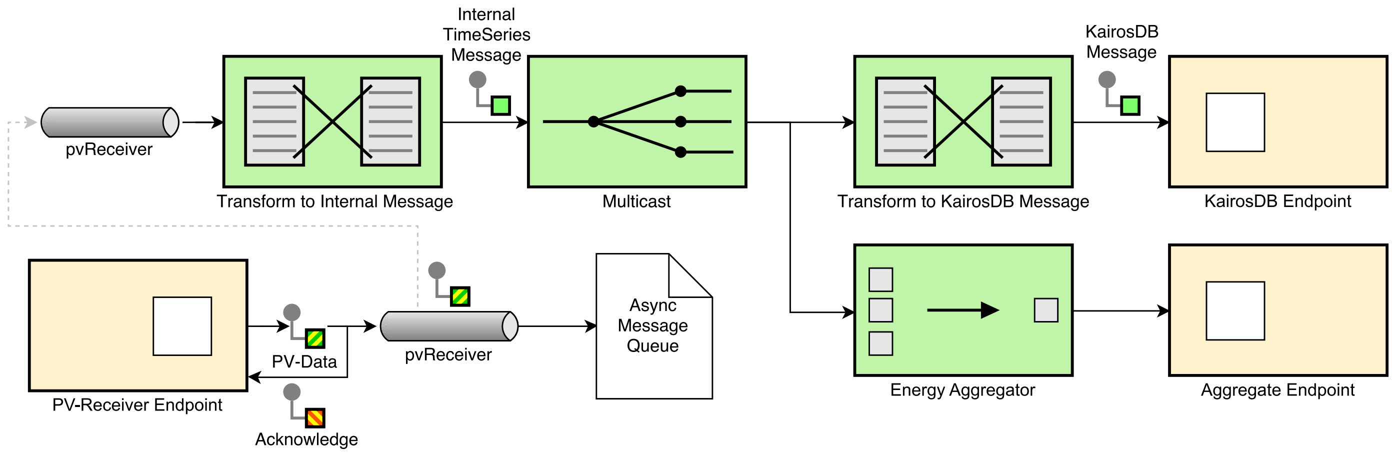


Figure 2. Visualisation of our general prototype based on EIP notation.

This tuple corresponds to the structure of data that are sent to a KairosDB instance for storing also. The included ID is not needed for the process of storing the data but necessary for matching the messages afterwards. Finally, the aggregation endpoint of this test setup gets its message as the same structured data like the data source tuple "(time, energy, station, id)".

The internal message routing has to be implemented across Node-RED, Apache NiFi and Apache Camel as shown in Figure 2. This figure illustrates the test setup and its components by using the notation of EIPs. First, there is an endpoint called "pvReceiver", which receives incoming messages and pushes them to a transformation step using an asynchronous message queue. Out of that, the endpoint ensures that the sender gets an acknowledge message. After transforming the information into an internal format, a multicast happens. So, one incoming message is sent to two other endpoints. These are the aggregate and the database endpoint. Before the data is sent, it is transformed to a matching database message format ("Transform to KairosDB Message" in Fig. 2) or the information is aggregated (("Energy Aggregator" in Fig. 2)).

The selected transformation and routing steps refer to the already mentioned requirements in Section IV and architectural draft in Section III to cover some data source, transformation, processing, reverse transformation as well as dumping.

A. Node-Red

Node-RED is a JavaScript-based message processing framework with IoT roots and can be used to solve application integration problems quickly. The framework is executed with Node.js and uses NPM for dependency management. Implementing the test setup mentioned above within Node-RED web client can be done by using a bunch of function nodes, nodes to create HTTP endpoints as well as change nodes. Change nodes are designed to modify the structure of our currently handled message object. Function nodes, on the other hand, are designed to execute custom scripts onto a particular message. Finally, there are nodes to create HTTP endpoints. Examples are HTTP server nodes to some path which can be called, HTTP response nodes which have to be placed within a message processing path which starts with an HTTP server node and HTTP client nodes to call external resources.

The implemented setup is shown in Figure 3. As mentioned, the messaging pipe starts with "PV Receiver" to create an HTTP server endpoint for "/endpoints/pvenergy". The message is piped onto an HTTP response node as well as to the primary processing path. The path starts with a function node to clean, enrich and transform incoming messages into the internal format. The result is forwarded to the database handling as well as the aggregation processing. Our database handling creates KairosDB compatible messages by using a template node and submits the resulting message by using an HTTP client node. The aggregation processing utilises the other function node to implement the aggregation function. This function node describes a simple memory to persist messages within a time window of ten seconds as well as calculating the mean within this window for the particular installation. The aggregation handling is finalised with a switch node to determine "NaN" values and an HTTP client node.

Summarising, Node-RED is a platform which is quickly providable for fast prototyping which can integrate various data sources as well as data sinks. But, it is tricky to develop collaboratively. Well, each developer can maintain its environment, but Node-RED itself manages Node-RED-Flows; synchronising them between different development platforms is hard. Furthermore, any particular use case, e. g., aggregating values from messages has to be implemented manually or by using additional NPM-based components which can be added directly in Node-RED. However, it is possible to integrate a broad range of endpoints with standardised formats and protocols. Handling proprietary endpoints requires more efforts in development.

B. Apache NiFi

Apache NiFi is a tool that runs within a Java Virtual Machine (JVM). A graphical user interface is offered within the web browser. Multiple of the so-called "Processors" can be used for standard tasks like receiving and sending HTTP requests ("PVReceiver" or "PostToKairos" in Fig. 4). Out of that, one can use custom processors by providing external JavaScript files (transform nodes in Fig. 4) or external java packages ("EnergyAggregator" in Fig. 4).

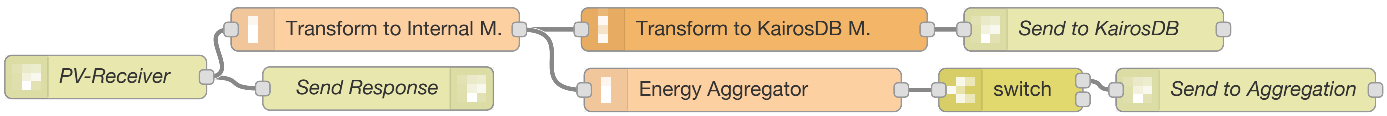


Figure 3. Node-RED implementation of the example from Section VI.

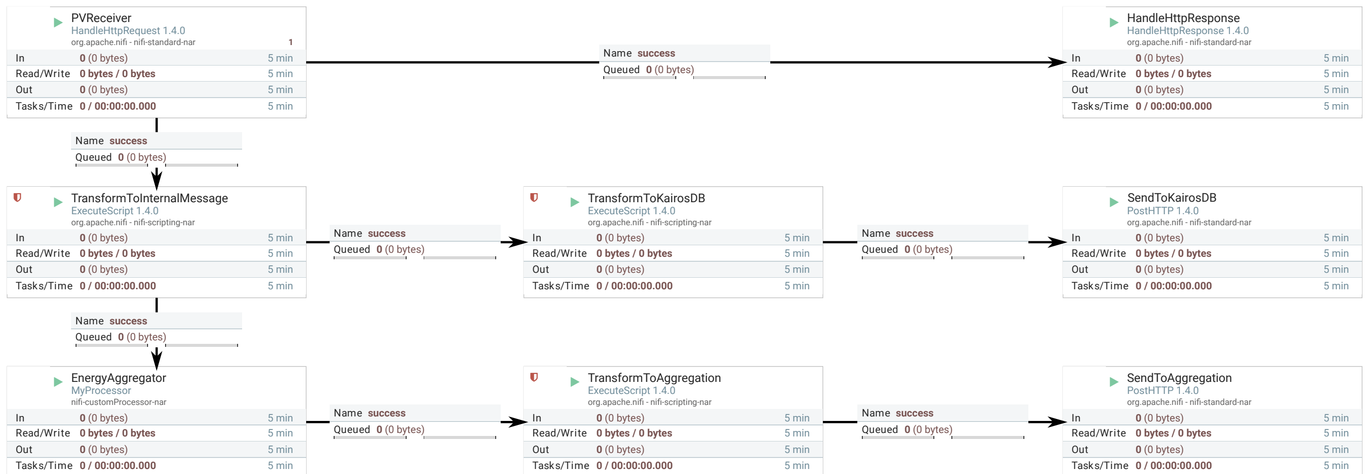


Figure 4. Message processing using Apache NiFi.

The “PVReceiver” accepts JSON-formatted data sent to the matching internet protocol (IP) address on port 8000. The processors “SendToKairos” send two different message formats to two different destinations (“http://nifireceiver:1880/dbmessages” and “http://nifireceiver:1880/averagemessages”). After receiving a message via POST request, it is answered by the “HandleHttpResponse” processor. The message is transformed into an internal format and processed on two different paths. The upper one (Fig. 4) transforms the message to another format and sends it to a Kairos endpoint. The lower branch aggregates the energy values of the messages in a way as mentioned above and sends them to the database as well. In contrast to our general setup (Fig. 2), we used an additional formatting step before publishing to the aggregation endpoint. So, we were able to separate the calculation from the formatting step.

We have to mention some application-specific facts. The incoming information is distributed within Apache NiFi by using so-called “Flowfiles”. These contain attributes added by the processors like HTTP header data. Out of that, the user can add attributes to custom scripts. So, we use the attributes to map the energy and time values that should be processed within our use case. In the end, the “TransformToKairos” processors (Fig. 4) take the matching attribute values and put them into the outgoing message. The “EnergyAggregator” (Fig. 4) internally uses a map to calculate station-specific averages over the last ten seconds. Thus, a list of measurement values and timestamps is managed for each station. Old values are removed from the list, so each calculation happens on the actual values.

We configured the “PVReceiver” with an internal queue size of 1000 requests. Out of that, every queue between the processors is configured with a “maxWorkQueueSize” of 10^6 and a “maxWorkQueueDataSize” of 1 GB. If we had not done that, overfull queues would cause the preceding processors to pause their work. In theory, this “maxWorkQueueSize” enables Apache NiFi to keep all incoming messages within one single queue.

Finally, we can say, that Apache NiFi provides a nice workflow for creating custom processors and integrating own functions. JavaScript can be used via external script files. Furthermore, a Maven template can be used to create processors using Java. Every manipulation of the data flow graph within the web browser causes a history file, which is stored in an archive directory. So, a rollback can be done quickly. We want to outline, that there were no problems of dockerizing this application. It is easy to provide the script files and processor archives by using volumes and corresponding configuration files.

C. Apache Camel and Wildfly

Apache Camel is a Java-based EAI-framework, which is lightweight and extendable. It can be executed as a standalone routing system or within middleware infrastructures like Spring, Java EE, Apache ServiceMix or JBoss Fuse. Implementing the test setup mentioned above within Apache Camel can be done by utilising a REST endpoint and describing a route which channels incoming messages to our HTTP database and aggregation endpoints. Apache Camel offers a large number of implemented patterns, which are described in [6], as well as the option to implement custom processes, for example within “Beans”. Furthermore, it is possible to extend the framework with own components for

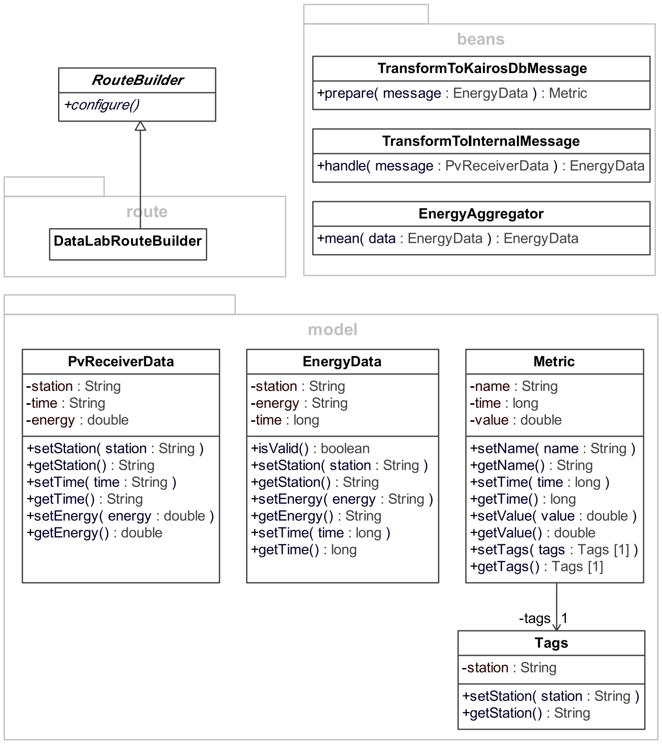


Figure 5. Apache Camel implementation of the example from Section VI.

further functionalities.

Fig. 2 visualises general and the finally implemented route within Apache Camel. Its components are shown in Fig. 5. The route itself is implemented by using the so-called “Java Domain Specific Language (Java DSL)” in Apache Camel. This route is implemented within “DataLabRouteBuilder” and describes the REST endpoint, which uses a servlet to process a specific resource and utilises SEDA to decouple incoming message flows from database and aggregation flows locally. SEDA is a lightweight in-memory message queue component within Apache Camel. The decoupled route contains the transformation and enrich bean “EnrichPvReceiverData” to transform external “PvReceiverData” into internal “EnergyData” as well as a multicast to handle the database and aggregation route. The database route contains another bean “KairosDbPrepare” to transform internal “EnergyData” into “Metric” datatypes for “KairosDb”. The aggregation route includes the aggregation bean “AggregationByInstallation” itself, which is implemented as stateful bean to save messages within a time window of ten seconds and finally calculate the mean for a particular installation. Both routes are completed with an HTTP client call onto the respective external endpoint.

Finally, Apache Camel is easy to use, primarily when used in combination with Maven as build and deployment tool. It is possible to describe routes within Java DSL, as we did, or use XML-based description to build those routes. Furthermore, Apache Camel is primarily a routing engine. Any particular use case, e. g., aggregating values over messages, has to be implemented manually or by using additional libraries.

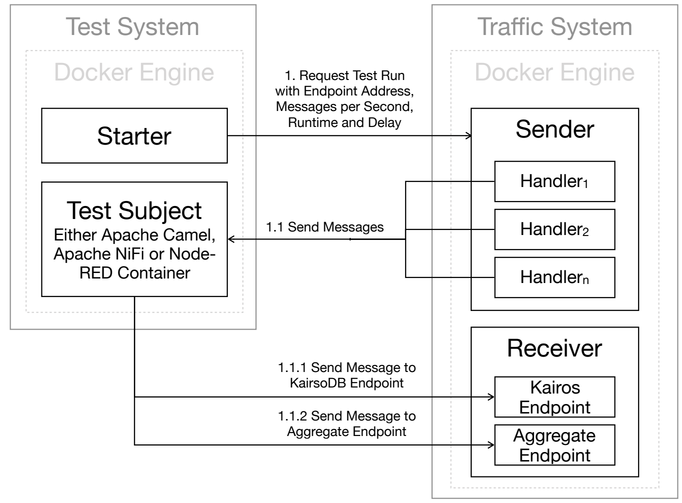


Figure 6. Visualisation of the test setup and the resulting communication steps.

D. Comparison

Table II compares our three prototypes. The characteristics regard the type of development, how to configure the data flow, the programming language for development, available scripting languages, possibilities for extension, resulting artefacts and containerization abilities were compared. Differences between the tools and the prototypes result primarily in the type of development, which is either based on source code (Apache Camel) or a web-based GUI (Apache NiFi and Node-RED), as well as in the resulting artefacts. Apache Camel is packaged and run as a traditional JAR archive; thus, it can run in execution environments with installed Java. In the case of Apache NiFi and Node-RED, however, the deployment mainly revolves around the description of the processing (flow file) and the associated dependencies or any self-implemented extensions.

VII. Runtime Measurements

In this section, we want to test the prototypes mentioned above. Guaranteeing constant conditions for every application and run, Docker containers are executed on the same machine. These containers encapsulate the runtime environment as well as the prototype itself. Our test machine runs on Debian GNU/Linux 9.3 Stretch using an Intel(R) Core(TM) i5-4570 CPU @ 3.20GHz. We chose this machine because we target to reach the hardware limitations faster. Because of the main task of our prototypes is routing messages, some exclusions are necessary. First, the application sending information to the routing engine is installed on another machine. Furthermore, the service which receives information sent by the routing engine is placed on another machine too. This setup admits for quantifying the response time, memory consumption and CPU load of the various Docker containers or the applications within them omitting the aspect of additional load of sending and receiving applications.

The setup used for the test execution can be found under [40] as a source code repository and is visualised in Fig. 6. Within the repository, you can find various docker-compose files [41], which start the different tools and a so-called

TABLE II. Prototype comparison.

Characteristic	Apache Camel	Apache NiFi	Node-RED
Development	Source code	Web-based GUI	Web-based GUI
Configuration	Possible to include external configuration, default flow described and configurated in DSL with Java or XML	Embedded in flow, has to be edited by using GUI	Embedded in flow, has to be edited by using GUI
Programming Language	Java	Java	JavaScript
Scripting	-	Clojure, ECMAScript, Groovy, Lua, Python, Ruby	JavaScript
Extending	API available, has to deployed as JAR and included in classpath, also possible to use simple Java Beans without any API dependency	API available, deployed as JAR and has to be placed in classpath	API available, installed via NPM
Artifacts	Packed as JAR archive	Flow file and dependency JARs	Flow file and dependency list
Containerise	Common workflow with Maven and Docker, use Java runtime environment	Use Docker, add flow and dependency JARs to existing Apache NiFi container	Use Docker, add flow and dependency list to existing Node-RED container

"Starter Container". This one sends a message to the sender, transmitting the destination IP address of the tool container, a time delay and the duration of the sending process. Afterwards, the starter container is powered off. Using this approach, the tool to evaluate and the test configuration can take place on the test machine. The measurement process is started by it too, and a script automates the whole process.

The sending device transmits JSON-based structured data tuples like "(time, energy, station, id)". One could think of a solar power system with a particular station identifier sending the actual energy production. The frequency of transmitted messages is configurable and initially set to 200 per second. Later we increased the rate up to 800 messages. This procedure supplements the measurements of the formerly published paper [1]. The sending process itself had a duration of 600 seconds. After this sending time, we waited for 600 seconds for later arriving responses. So, our measurement phases took 1200 seconds overall.

The JVM for our Apache NiFi CEP instance and the Apache Camel prototype are fixed to 6 GB of space, which is fully allocated on startup. The "MaxPermSize" is set to 1 GB. We use the measured values of "jstat" for calculating the memory consumption of the tools mentioned above with a time resolution of one second. Furthermore, we sum up the usage of survival space ("S0U" and "S1U"), eden space ("EU"), old space ("OU"), metaspace space ("mu"), and compressed class space ("ccsu"). A node.js module measures the memory consumption of Node-RED, i. e., the "heapUsed" value. Out of that, the CPU load is measured by the "top" command every second. We get the response times of the various systems by measuring the time of sending and the time of receiving messages in milliseconds. The arrival timestamps of messages corresponding to database operations and aggregations are measured separately.

The following results are presented first in an error-specific manner. We refer to individual behaviour and found error cases. Afterwards, the tools are compared to each other.

A. Results

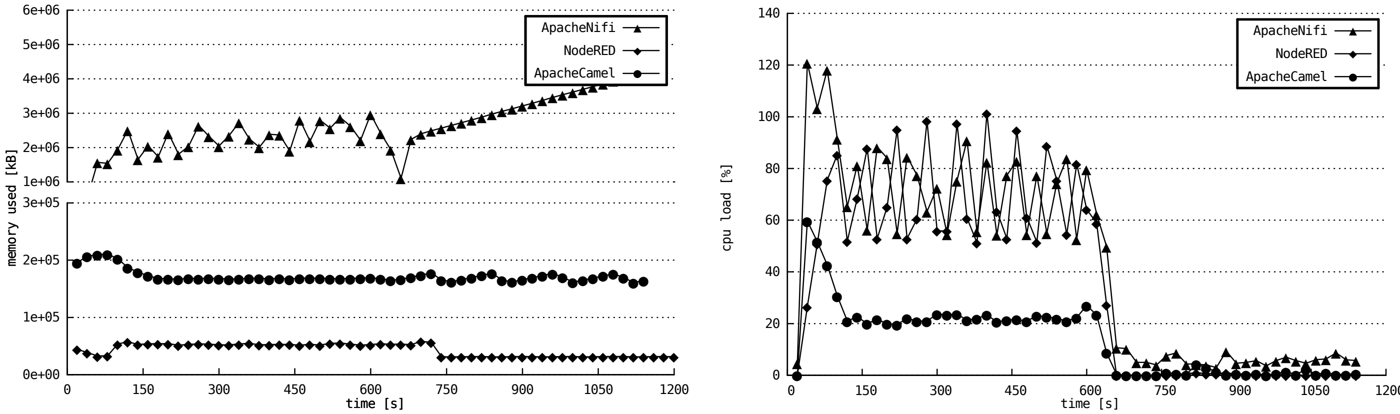
By sending 200 messages per second, we did not find any errors in message processing. Every tool did its part and processed all messages we sent. On higher message rates Node-RED did not act like the other tools. Beginning with a rate of 400 messages per second it did not answer the requests appropriately. About 602 of 239601 messages we tried to send were not answered, i. e., our sender delivered errors. The rate of 800 messages per second caused about 226,303 TCP-related errors, like "EADDRNOTAVAIL", "ECONNRESET" or "ETIMEDOUT" while 480,001 messages were sent overall. For this high message rate, we can state, that Node-RED was able to process only the half of the messages correctly. Apache NiFi shows another behaviour. Sending 800 messages to it, about 63,000 messages of 479,201 cannot be processed. In contrast to Node-RED, we got the HTTP error 503 for a not available service, i. e., the message could not be processed at all, database and aggregation message. Apache Camel always answered all messages send to it.

We were caused by the error messages to research more intensively on the mistakes happen. We can state that the TCP handshake for Node-RED did not happen in the right way. We checked the amount of opened ports of the involved Docker containers and the machines, but there is no lack. We state, that the Javascript event processing loop itself needs that much CPU and memory, that not all requests of the sender get handled quick enough. So, many messages the sender tries to dispatch are not delivered to the processing parts within Node-RED.

Summarizing we can say, that our measurements got difficult because of instabilities of the Docker Engine. Further analyses may use a more clean Docker setup, i. e., shutting down the service or restarting the test machine before each series of measures.

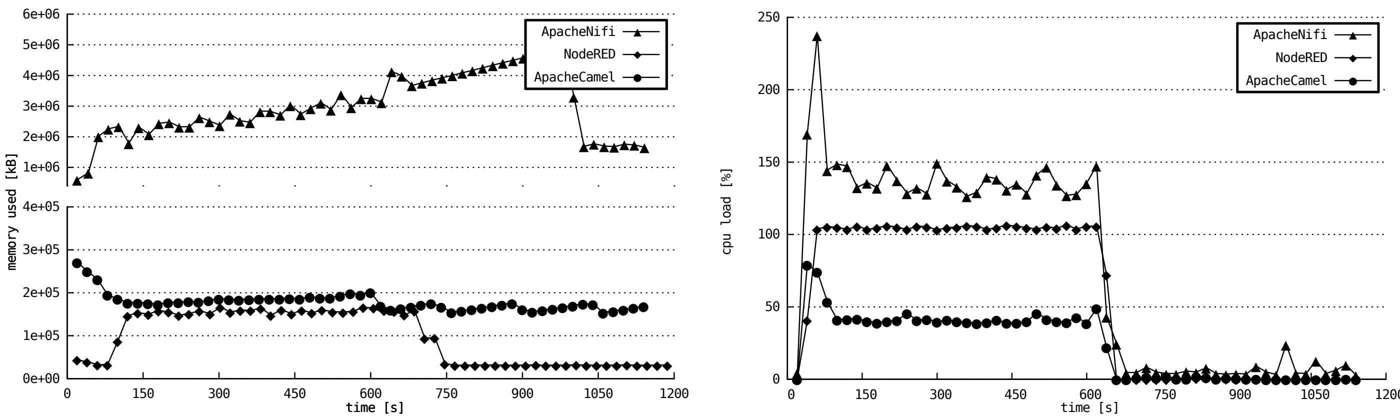
B. Comparison

The memory consumption (Fig. 7(a) and Fig. 9(a)) does not change much across all measurements (200 up to 800 messages per second). A factor of 10 is between the memory consumption of Apache Camel or Node-RED and Apache NiFi. The "S1U" value measured by "jstat" is the main cause of the higher RAM consumption of Apache NiFi. We want to mention the decreasing memory usage of Apache Camel



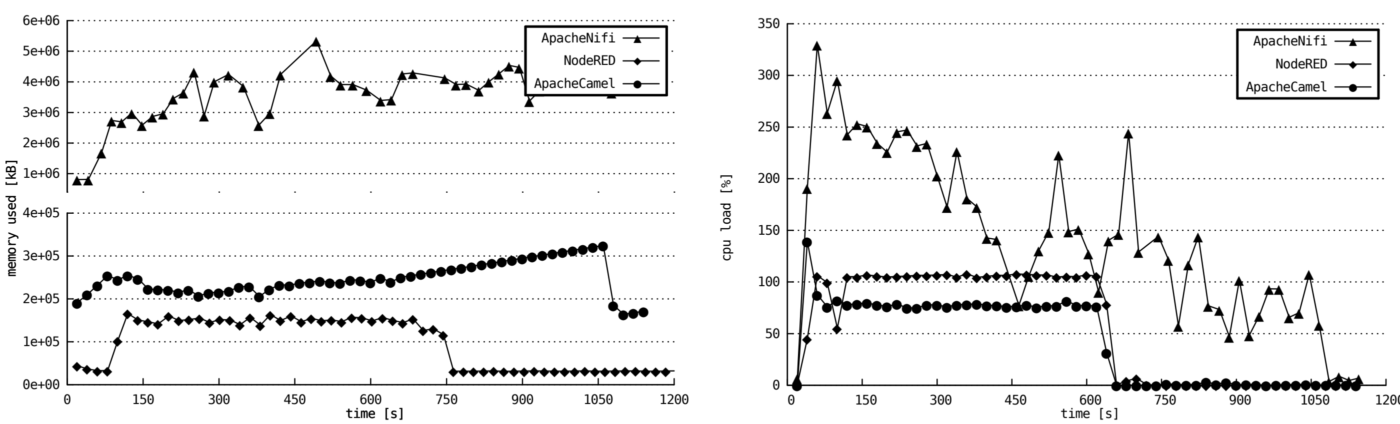
(a) Memory Consumption (b) CPU Usage

Figure 7. Memory consumption and CPU load while sending 200 messages per second (values aggregated over 20 s).



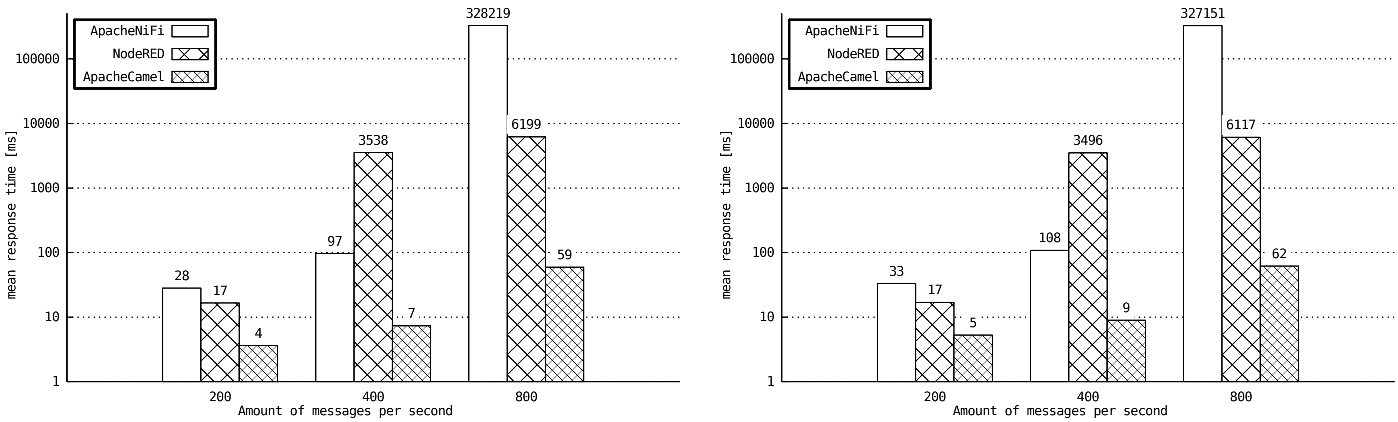
(a) Memory Consumption (b) CPU Usage

Figure 8. Memory consumption and CPU load while sending 400 messages per second (values aggregated over 20 s).



(a) Memory Consumption (b) CPU Usage

Figure 9. Memory consumption and CPU load while sending 800 messages per second (values aggregated over 20 s).



(a) KairosDB Endpoint

(b) Aggregate Endpoint

Figure 10. Mean response times of tested systems with various message frequencies for database (a) and aggregation (b) messages.

while processing messages. According to the former published paper [1] we state, that "eden space utilisation" causes the less memory consumption while processing a more significant amount of messages. This finding is also present this time.

Watching the CPU graph (Fig. 9(b)) we can state, that Apache NiFi forces a much higher load than the other tools. Apache Camel showed an expectable process from a CPU usage of approximately 22 % for 200 messages per second up to 40 % for 400 messages per second and 80 % for the rate of 800 messages. In contrast, Node-RED reached the limit of a CPU usage of 100 % by processing 400 messages per second (Fig. 8(b)) and kept this behaviour on 800 messages. Apache NiFi showed the highest usages of the CPU. At this point, we want to mention, that the higher CPU usages (more than 100 %) in Fig. 7(b) and Fig. 9(b) are caused by the usage of Java. Apache Camel and Apache NiFi are able to use multiple threads so that the sum of CPU usages can sum up to more than 100 %. Node-RED, in contrast, is limited to one single thread, as we would expect for a Nodejs application.

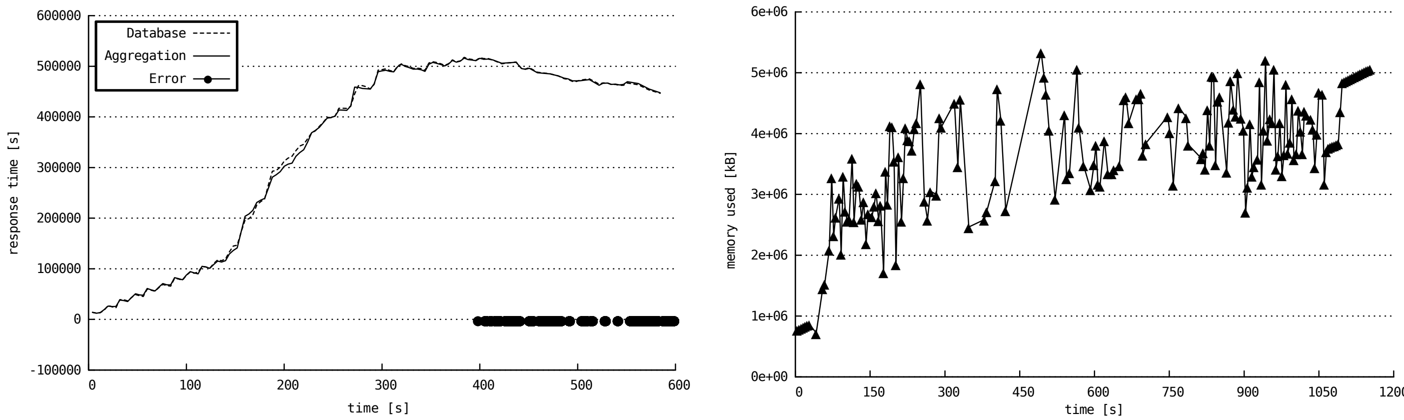
We researched more intensively on the behaviour of Apache NiFi. As visualised in Fig. 11(a), the response time increases rapidly after 150 seconds. The step pattern up to this point in time may cause by small garbage collecting processes, which clean the memory from processed message objects. At this time, Fig. 11(b) shows a nearly full available RAM. So, from this time Apache NiFi needs much time to process the incoming messages, because of a quiet full memory and the corresponding more comprehensive garbage collections. Out of that, shown in Fig. 11(a) on the lower right (plotted as negative values), the HTTP endpoint delivers errors to the sending component. The corresponding points in Fig. 11(b) show a drop in memory consumption. We state that Apache NiFi prevents its processors from overfilling the memory, even if not all messages can be received. Finally, it can see that the figures in Fig. 11 both correspond to each other. While we can see the ordinary memory behaviour of a java program on the right after 1080 seconds, we see the highest response times on the left at approximately 410 seconds. This timespan is approximately the time after the end of sending (600 seconds) when the memory consumption takes a regular course.

Other essential measurement values are the response times of the various tools, i. e., the time between sending a message and getting the calculated result for aggregation or the reformatted database message. As mentioned above (Section VII-A) not all messages were answered. We calculated the mean response times overall sent and received messages. So, Fig. 10 must be considered from the point of view that Apache NiFi was not able to process about one-eighth of the messages (63357 of 479201) at the highest rate. Out of that, Node-RED caused errors for the half of the messages (226734 of 480001) the sender tried to transmit. The results show quite similar times for database (Fig. 10(a)) and aggregation messages (Fig. 10(b)). However, especially while watching the logarithmic scale, we can state, that Apache Camel answers our requests very fast, even for high loads. Apache NiFi needs more time for high message rates, i. e., four orders of magnitude, but answers the requests in case of a not available service with a well-formed error code. Node-RED causes errors in the sending process and needs a higher answer time (two orders of magnitude) than Apache Camel.

We tested Apache NiFi in an additional setup. The incoming messages had to be passed through without any manipulation. The tool only had to accept a request and send the message content to an HTTP interface. The measured response times were in the range of 9 milliseconds (200 messages per second) up to 390 milliseconds (800 messages per second) by processing every message correctly. Even without aggregating or reformatting any message, Apache NiFi is slightly slower than Apache Camel.

VIII. Discussion

Within this section, we do not discuss the measurements, but the process of measurement itself. Some peculiarities were referring to, e. g., the network connection. At first, we have to state some problems referring to our measurement setup. We used a common SoHo-router for our initial test setup. But, sending 800 messages per second caused network problems of the router itself, so a complete reboot was necessary. Afterwards, we changed our network setup to a common network switch. The sender and the test machine were configured with static IP addresses. We got a more stable network infrastructure in that way. Overcoming this problem,



(a) Response Time (b) Memory Consumption

Figure 11. Response times of Apache NiFi along the sending process of 800 messages per second (a) (3200 values aggregated, i. e., four seconds) and the corresponding memory consumption of Apache NiFi (b) (values aggregated over 4 s).

we found, that our sending node did not work correctly for high message rates (800 messages per second), at least in combination with the Node-RED test setup. Not all messages we prepared for sending produced log entries. Missing TCP responses caused this, the sender did not get them. So, we restructured our sending process to use multiple threads and to produce more log entries, i. e., for creating a message, starting of sending process and logging errors. Referring to Apache NiFi, we may decrease its response time by using custom processors for message formatting instead of JavaScript using script executors. In total, to say is that a message rate of 800 ones per second forces our test setup. Especially Fig. 9(a) shows irregular time distances of the Apache NiFi data. This irregularity is caused by lags of the measurement tool itself, which is created using Java. Furthermore, the docker daemon on the test machine itself caused an unidentifiable error, which forced us to reboot the Docker Engine. We were not able to call commands within our test container or copy the locally created measurement files for memory and CPU usage. Our assumption referring to this is a broken docker process. For further measurements, we should think of regular reboots. Maybe a process controlled by Wake-On-LAN is more suitable for this use case because we would restart the entire machine instead of single processes, that can cause errors.

IX. Conclusion

The WDL has to be able to handle data streams as mentioned in different manners. Beside the integration and routing itself, there are tasks in the area of complex event processing as well as knowledge discovery in data. This is our second reflection of architectural backbone technologies covering those aspects. We tested high message rates forcing the test machine up to the limit of the hardware. Based on our experiences and measurements gathered from this test setup, we can make some decisions. In the case of a complex heterogeneous environment with different kinds of interfaces, Apache Camel seems to be a right choice. It is used within a wide range of conditions and able to handle many technologies to cover integration problems. Furthermore, this tool can manage high message rates by using reasonable memory.

Node-RED is a well-suited tool for rapid prototyping and IoT. On higher loads, it causes errors that are hard to handle. So, a conceivable approach could be inventing message processing using Node-RED at first. Afterwards, a more efficient implementation could be done using tools like Apache NiFi or Apache Camel. Node-RED might be usable as front-end system to easily integrate standardised external interfaces as well as an additional platform for experiments within a productive setup. Nevertheless, everything which can be done with Node-RED seems to be possible with Apache Camel too. The main difference can be found in the usability, the deployment process and the underlying language. Adapting knowledge discovery in such setups, independent of which routing engine is used, should be possible by using a database and route messages as required or by integrating available public interfaces from tools for knowledge discovery within Apache Camel or Node-RED. Apache NiFi shows a stable behaviour, even in case of high loads. If the processing of a message cannot be guaranteed, we get an HTTP error code and can try to send the information later. Furthermore, it seems to be quickly integrable with Apache ZooKeeper to run it within a cluster. Such a setup may be possible with Apache Camel or Node-RED.

Further research could be done on the possibility of using the considered tools within a clustered environment. This environment could overcome load peaks and increase the availability of the system. Additionally, topics regarding security and privacy should be taken into consideration.

X. Acknowledgements

The research project WINNER is funded by the Federal Ministry for Economic Affairs and Energy of Germany under project number 01ME16002D.

References

[1] Sebastian Apel, Florian Hertrampf, and Steffen Späthe. Evaluation of architectural backbone technologies for winner datalab. In *Proceedings of the Sixth International Conference on Smart Cities, Systems, Devices and Technologies*, pages 35–43, Venice, June 2017.

[2] Chemnitzer Siedlungsgemeinschaft eG. WINNER-Projekt, 2017. Website http://www.winner-projekt.de; 2018-02-01.
[3] D.S. Linthicum. *Enterprise Application Integration*. Addison-Wesley information technology series. Addison-Wesley, 2000. ISBN 9780201615838.
[4] David Chappell. *Enterprise Service Bus*. O'Reilly, 2004.
[5] Falko Menge. Enterprise service bus. In *Free and Open Source Software Conference*, 2007.
[6] Gregor Holpe. Enterprise integration patterns. In *Proceedings of 9th Conference on Pattern Language of Programs*, September 2002.
[7] Michael Eckert and François Bry. Complex event processing (cep). *Informatik Spektrum*, 32(2):163–167, 2009.
[8] Gianpaolo Cugola and Alessandro Margara. Processing flows of information: From datastream to complex event processing. *ACM Computing Surveys*, 44(3):1–70, 2012.
[9] Andrew G. Psaltis. *Streaming Data - Understanding the real-time pipeline*. Manning Publications Co., 2017.
[10] Christian Gottermeier. Data mining: Modellierung, methodik und durchführung ausgewählter fallstudien mit dem sas enterprise miner. Diplomarbeit, Universität Heidelberg, 2003.
[11] Ann Chervenak, Ian Foster, Carl Kesselman, Charles Salisbury, and Steven Tuecke. The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. *Journal of Network and Computer Applications*, 23(3):187 – 200, 2000. ISSN 1084-8045. doi: 10.1006/jnca.2000.0110.
[12] B. Tierney, R. Aydt, D. Gunter, W. Smith, M. Swany, V. Taylor, and R. Wolski. A grid monitoring architecture, 2002.
[13] Hans-Jürgen Appelrath, Petra Beenken, Ludger Bischofs, and Mathias Uslar. *IT-Architekturentwicklung im Smart Grid*. Springer Gabler, Heidelberg, Germany, 2012.
[14] S. Rusitschka, K. Eger, and C. Gerdes. Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain. In *2010 First IEEE International Conference on Smart Grid Communications*, pages 483–488, Oct 2010. doi: 10.1109/SMARTGRID.2010.5622089.
[15] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash. Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys Tutorials*, 17(4):2347–2376, Fourthquarter 2015. ISSN 1553-877X. doi: 10.1109/COMST.2015.2444095.
[16] G. Choudhary and A. K. Jain. Internet of things: A survey on architecture, technologies, protocols and challenges. In *2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–8, Dec 2016. doi: 10.1109/ICRAIE.2016.7939537.
[17] Miao Wu, Ting-Jie Lu, Fei-Yang Ling, Jing Sun, and Hui-Ying Du. Research on the architecture of internet of things. In *2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)*, volume 5, pages V5–484–V5–487, Aug 2010. doi: 10.1109/ICACTE.2010.5579493.
[18] Zhihong Yang, Yingzhao Yue, Yu Yang, Yufeng Peng, Xiaobo Wang, and Wenji Liu. Study and application on the architecture and key technologies for iot. In *2011 International Conference on Multimedia Technology*, pages 747–751, July 2011. doi: 10.1109/ICMT.2011.6002149.
[19] A. Krylovskiy, M. Jahn, and E. Patti. Designing a smart city internet of things platform with microservice architecture. In *2015 3rd International Conference on Future Internet of Things and Cloud*, pages 25–30, Aug 2015. doi: 10.1109/FiCloud.2015.55.
[20] Sam Newman. *Building Microservices*. O'Reilly, 2015.
[21] Jean-Louis Maréchaux. Combining service-oriented architecture and event-driven architecture using an enterprise service bus. *IBM Developer Works*, pages 1269–1275, 2006.
[22] M. Vianden, H. Lichter, and A. Steffens. Towards a maintainable federalist enterprise measurement infrastructure. In *2013 Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement*, pages 63–70, Oct 2013. doi: 10.1109/IWSM-Mensura.2013.20.
[23] M. Vianden, H. Lichter, and A. Steffens. Experience on a microservice-based reference architecture for measurement systems. In *2014 21st Asia-Pacific Software Engineering Conference*, volume 1, pages 183–190, Dec 2014. doi: 10.1109/APSEC.2014.37.
[24] Climate Data Center. Website http://www.dwd.de/DE/klimaumwelt/cdc/cdc_node.html; 2017-01-06.
[25] OpenWeatherMap, 2017. Website http://openweathermap.org/price; 2017-01-06.
[26] Apache Camel, 2015. Website http://camel.apache.org; 2016-06-22.
[27] Apache Storm, 2015. Website http://storm.apache.org; 2017-02-28.
[28] Apache Spark, 2015. Website http://spark.apache.org; 2017-02-28.
[29] Apache Hadoop, 2014. Website http://hadoop.apache.org; 2017-02-28.
[30] Apache ServiceMix, 2011. Website http://servicemix.apache.org; 2017-02-28.
[31] Jboss Fuse, 2016. Website https://developers.redhat.com/products/fuse/overview/; 2017-02-28.
[32] Apache NiFi, 2018. Website https://nifi.apache.org/; 2018-02-19.
[33] Siddhi Complex Event Processing Engine, 2017. Website https://github.com/wso2/siddhi; 2017-03-05.
[34] Esper, 2016. Website http://www.espertech.com/esper/; 2017-03-01.
[35] WSO2, 2017. Website http://wso2.com/products/complex-event-processor/; 2017-01-06.
[36] RapidMiner, 2017. Website https://rapidminer.com/; 2017-03-05.
[37] KNIME, 2017. Website https://www.knime.org; 2017-03-05.
[38] Node-RED, 2017. Website https://nodered.org; 2017-01-06.
[39] StreamLine - Streaming Analytics, 2017. Website https://github.com/hortonworks/streamline/; 2018-02-28.
[40] Architectural backbone evaluation source code, 2018. Website https://github.com/winner-potential/smart-2017; 2018-02-28.
[41] Docker, 2017. Website https://www.docker.com/; 2017-12-05.

Adaptive Portfolio Asset Allocation Optimization with Deep Learning

Samer Obeidat
Investifai Inc.
Ottawa, Ontario, Canada
Email: sam@investifai.com

Daniel Shapiro
School of Electrical Engineering and Computer Science
University of Ottawa
Ottawa, Ontario, Canada
Lemay Solutions Consulting Inc.
Ottawa, Ontario, Canada
Investifai Inc.
Ottawa, Ontario, Canada
Email: daniel@investifai.com

Mathieu Lemay
Lemay Solutions Consulting Inc.
Ottawa, Ontario, Canada
Email: matt@lemay.ai

Mary Kate MacPherson
Lemay Solutions Consulting Inc.
Ottawa, Ontario, Canada
Email: marykate@lemay.ai

Miodrag Bolic
School of Electrical Engineering and Computer Science
University of Ottawa
Ottawa, Ontario, Canada
Email: mbolic@eecs.uottawa.ca

***Abstract*—Portfolio management is a well-known multi-factor optimization problem facing investment advisors. The system described in this work can assist in automating portfolio management, and improving risk-adjusted returns. The asset allocation action recommendations were personalized to the portfolio under consideration, and were examined empirically in this work in comparison to standard portfolio management techniques. This work presents a Long Short-Term Memory approach to adaptive asset allocation, building upon prior work on training neural networks to model causality. The neural network model discussed in this work ingests historical price data and ingests macroeconomic data and market indicators using Principal Components Analysis. The model then estimates the expected return, volatility, and correlation for the selected assets. These neural network outputs were then turned into action recommendations using a Mean-Variance Optimization framework augmented to use a forward-looking rolling window technique. Testing was performed on a dataset with a 7.66 year duration. The observed mean annualized return for classical passive portfolio management approaches were 4.67%, 3.49%, and 4.57%, with mean Sharpe ratios of 0.46, 0.20, and 0.54. 10 simulations using the new Long Short-Term Memory model from this work provided a mean annualized return of 10.07%, with a Sharpe ratio of 0.98. This work provides the conclusion that a Long Short-Term Memory model can generate better risk-adjusted returns than conventional strategic passive portfolio management.**

Keywords–Recommender systems; Deep learning; Portfolio management.

I. Introduction

Whereas our prior work investigated learning causality from observing computer user actions [1], this extension of that work continues that line of research, using Long Short-Term Memory (LSTM) neural networks to model the cause and effect inherent in portfolio management decisions. The objective of this work is to develop an artificial intelligence (AI) approach for adaptive investment portfiolio management by examining the assets for a long time period and by looking and diverse asset classes on a global scale. This approach will be contrasted against conventional strategic passive investment portfolio management strategies. Adaptive asset allocation is a timely area of research, as deep learning innovations are being productized to create financial products such as robo-advisors [2], ETFs [3], and hedge funds [4]. The decision engine from [1] was replaced by the Markowitz's Mean-Variance Optimization (MVO) framework, and the learning algorithm was replaced by a LSTM model. Prior work on LSTM portfolio management [5] was extended in this work.

In recent years, there has been an increasing focus on investments into passively managed funds. These aim to replicate the market's performance rather than beating it [6]. In 2011, Burton Malkiel made the case for passive investing by empirically proving that investing in the S&P 500 during the period of 1969 to 2010 would have generated 80% more returns than the average actively managed fund [7]. This has encouraged investment managers to enjoy the benefits of diversification using cross-asset correlations within a certain risk profile [8]. The dilemma of managing a portfolio has thus become deciding what proportion of the investment should be allocated to each asset available. Modern portfolio theory (MPT) suggests the use of the MVO framework to eliminate risks that are not systematic, and determine the optimal distribution of the investment between the different assets. It is important to note here that the optimality of the asset allocation does not imply that the estimate of asset-related features is optimal. Rather, on the assumption that the asset-related parameters were error-free, the MVO asset allocation is guaranteed not to add any additional error.

Markowitz's MVO framework aims to achieve portfolio diversification while minimizing specific risks and determining

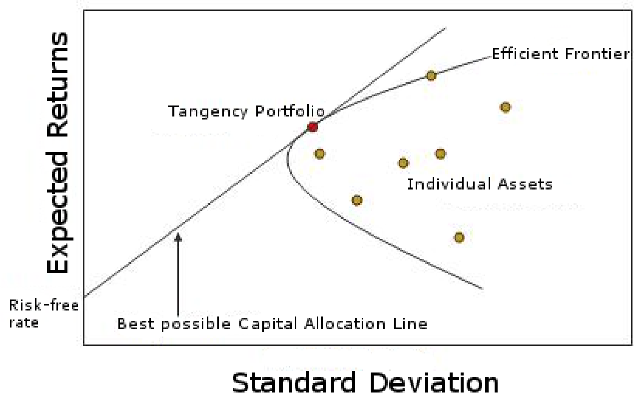


Figure 1. Markowitz Frontier (Modified from Wikimedia Commons) [12]

the risk-return tradeoffs for each asset [9]. The first step for MVO is to develop estimates of the expected returns and volatilities for each asset. These estimates can be achieved using classical indexing models such as the Capital Asset Pricing Model (CAPM) [10] or the Three-Factor Model (3FM) [11], or by using computational methods. After projecting the expected returns and volatilities for each asset, the algorithm selects the allocation of assets that has the highest expected return for a defined level of risk, or the lowest level of risk for a given level of expected return. As shown in Figure 1, the optimal portfolio lies on the hyperbolic curve called the efficient frontier.

This work examines the asset return and volatility estimates generated by LSTM Recurrent Neural Networks (RNNs). The model incorporates economic and market input features to adapt to changes in the financial markets, by regularly adjusting the asset allocation. This work used a globally diversified multi-asset portfolio consisting of 11 asset classes in a variety of different markets. Most of the previous studies have used short testing periods, while this work examined a broader testing horizon (January 2010 to August 2017).

The challenge of this work was to:

(TASK 1) Predict the return of any asset under consideration, by training an LSTM to model the asset performance.

(TASK 2) Optimize the allocation of assets within a model portfolio using the MVO framework.

The following sections review the prior art (Section II), present the causality-learning LSTM model (Section III and asset allocation model (Section IV), followed by a description of the dataset (Section V), a description of the empirical experimentation (Section VI) and presentation of the results (Section VII), closing with conclusions and future work (Section VIII).

II. Related Work

Random chance and many hidden variables can influence asset returns. It is therefore challenging to estimate the future returns for any asset, which is required in order to perform MVO asset allocation. Asset returns can be affected by economic conditions, commodities prices and political events, along with many other factors [13]. While conventional models tend to assume that relationships are linear when forecasting future returns, many of the real-world cases in the financial markets are non-linear [14]. Artificial Neural Networks (ANN) are excellent approximators of non-linear functions, and so the use of these ANN models in computational finance research has persisted. It has been known for decades that multilayer feedforward neural networks are capable of approximating any measurable function to any desired degree of accuracy [15]. The difference between ANN and other approximation methods is that it uses one or more hidden layers to transform the input variables, using a transfer function to deal with nonlinear statistical functions [16]. ANNs can analyse huge quantities of data to recognize patterns and make sense of incomplete or noisy data, and therefore provide an excellent alternative to linear models for forecasting and estimating financial time-series [17] [16].

An important part of the portfolio management process is the method of performance evaluation. Many techniques and models have been developed to evaluate portfolio performance based upon the portfolio's return-risk characteristics. The Sharpe ratio is measured by dividing the difference between the portfolio's expected return and the risk-free interest rate by the portfolio's standard deviation. The ratio computes the excess returns per unit of total risk [18]. The Treynor ratio is measured by dividing the portfolio's excess returns by the portfolio's beta (systematic risk) [19]. Jensen's alpha (α_i) is the intercept of Jensen's excess return of the single-index model, and it measures the abnormal return over holding the investment portfolio of an index fund [20]. As the various methods of weighting the portfolios have different standard deviations, this work uses the Sharpe ratio to enable comparison between results.

A number of studies have been conducted to evaluate neural networks' ability to predict financial time-series. The work available in the literature tends to focus on a specific stock market over a relatively short testing period. Freitas et al (2001) focused on the Brazilian Stock Exchange (BSE) over a 21 week testing period using an autoregressive neural network [21]. Jang & Lai (1994) focused on the Taiwan Stock Exchange (TSE) over two years of testing with a DAS network [22]. Ko & Lin (2008) had the same focus, looking at the TSE over a two-year period with a multi-layer resource allocation neural network [23]. Raei (2006) conducted a comparitive empirical study to examine the ability of a four layer Perceptron network to beat a classical portfolio that comprised of stocks from the Tehran Stock Exchange over a period of 13 months [24]. Overall, the literature lacks a strong body of comparative empirical analysis to examine the ability of machine learning (specifically LSTM neural networks) to achieve better estimates of the returns than the classical indexing models within multi-asset investment portfolios. As well, very few of them involved the use of a wide range of assets that covered different classes on a global scale. The vast majority of published work focuses on a specific market (such as a US market or Australian market) or a particular asset class (such as stock market or fixed income market).

Portfolios with exposure to global equity markets generate better risk-adjusted returns than portfolios dominated with domestic equities [25]. Empirical studies have also shown that globally diversified portfolios containing both equities and bonds outperform portfolios of equities alone [26]. Therefore, bonds should play a vital role in the portfolio.

One of the commonly-used classes of ANNs is the Recurrent Neural Network (RNN). While feed-forward networks are designed to have no feedback loops [27], RNNs contain at least one directed cycle to create an internal memory. Long Short-Term Memory (LSTM) networks are a form of RNN designed to deal with modeling temporal sequences [28]. Because of this, LSTMs can be readily applied to financial time series. Lee & Yoo (2017) looked into using LSTMs to predict potential returns of a variety of investments. They decided against using MVO to find the optimal set of assets' weights because MVO makes the simplifying and often incorrect assumption that asset returns are normally distributed [29]. Their dataset consisted of 10 top stocks in terms of market value from the S&P500 from 2004 to 2016. This work aims to expand on the use of LSTMs for asset allocation by bringing in additional market and economic data, to improve the quality of the predictions. It will also focus more on the parameters of the LSTM model itself.

III. Overall AI Model

Figure 3 presents the building blocks of the adaptive asset allocation system discussed in this work. The output from the system is a set of asset weights within an investment portfolio. The iterative process of asset allocation begins with preparing the input dataset, including historical returns of the 11 assets (ETFs) in the portfolio, and additional economic and market data that may influence future returns for each asset.

Once the input data has been prepared, Principal Components Analysis (PCA) is used to reduce the dimensionality of the additional economic and market data down from 387 features. After dimensionality reduction, only key economic and market features are passed along. This information compression helps to reduce the number of dimensions without much loss of information, going from a sparse to a dense data representation [30].

The reduced market and economic data are then passed to an LSTM RNN, along with the historical prices for the assets. Using this data, the RNN produces a prediction of the assets' future returns.

Finally, these predictions are used by the MVO model to generate the optimal weighting of the assets.

Figure 2 shows the overall algorithm of the model. Constraints a and b are chosen depending on the level of risk desired for MVO.

Input: Economic data $h0$; Market data $h1$; Assets' historical data h_2; Minimum weight constraint $a := 0.05$; Maximum weight constraint $b := 0.35$; TIP minimum weight constraint $a_{TIP} := 0.0$

Output: Optimal set of assets' weights $optimal\_asset\_allocation$

$e_0 = merge(h_0, h_1)$

$e_1 = scale(e_0)$

$reduced = PCA(e_1)$

$prices = scale(h_2)$

$returns = LSTM\_predict(reduced, prices)$

$optimal\_asset\_allocation = MVO(returns, a, b, a_{TIP})$

Figure 2. Overall LSTM-based asset allocation algorithm.

IV. Mean-Variance Optimizer

The MVO framework maximizes returns for a certain level of risk, or minimizes risk for a given expected return. MVO requires the estimations of expected returns of all included assets, their standard deviations, and the variance-covariance (or correlation) matrix in order to find the optimal asset allocations and calculate a set of efficient portfolios. The MVO method was used in this work to generate three outputs of asset allocations: CAPM-based weights, 3FM-based weights, and LSTM-based weights.

Given n assets, MVO requires expected returns on each asset r_i, standard deviation of returns σ_i, and covariance $\sum$ to be used as inputs to generate an efficient frontier of optimal portfolios. The total portfolio return r_p can be solved with the equation

$$r_p = \omega^t r$$

where ω is an $n \times 1$ column vector of portfolio weights, ω^t is the transpose of the vector ω, and r is an $n \times 1$ column vector of assets' returns r_i. The total portfolio variance σ_p^2 can be solved with

$$\sigma_p^2 = \omega^t \sum \omega$$

The asset allocation optimization problem can then be reduced to the following:

$$\begin{aligned} \underset{\omega}{\text{minimize}} \quad & \omega^t C \omega \\ \text{subject to:} \quad & r_p = \omega^t r, \\ & \sum_{i=1}^{n} \omega_i = 1, \\ & \text{No short selling constraint, thus } \omega_i \geq 0, \\ & a < \omega_i < b \end{aligned}$$

where C is an $n \times n$ covariance matrix, r_p is the total portfolio return, a is the minimum weight constraint, and b is the maximum weight constraint. For this work, the maximum weight constraint was set arbitrarily to 35% for all of the portfolio assets. The minimum weight constraint was set arbitrarily to 5% for all portfolio assets except for TIP, which had a minimum constraint of 0%. Treasury Inflation Protected Securities (TIPs) link principal and coupon payments to the Consumer Price Index (CPI), to help protect investors from inflation [31]. They have no minimum weight constraint because they are inefficient for investors with moderate to high risk tolerance [32].

V. Data Description and Analysis

The dataset used to evaluate and compare approaches consisted of historical, economic, and market data, as well as an economic event calendar.

A. Input Data

The input data can be grouped into three categories: asset historical data, macroeconomic data, and market data. While the classical models used only the historical data, the LSTM model used all three types of input data.

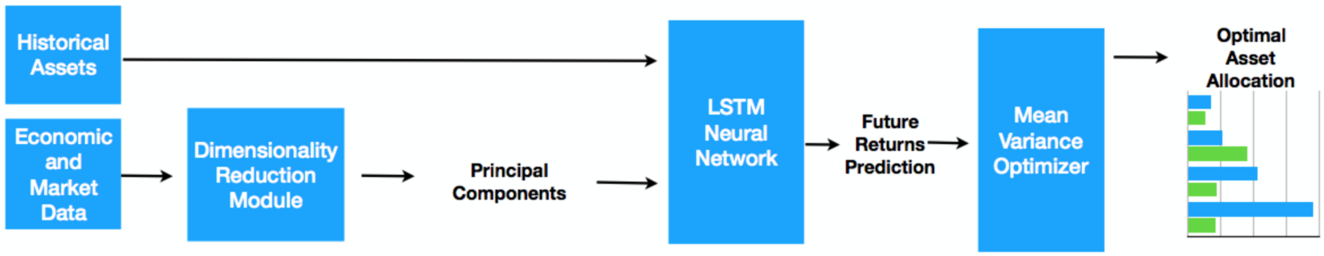


Figure 3. Block diagram for the overall asset allocation AI system [5].

1) Historical Data: Historical monthly adjusted closing prices were used to calculate the monthly expected returns and volatilities of each asset in the portfolio. The researched period was 115 months, from February 2008 to August 2017. The time-series data for each asset was obtained from Thomson Reuters (2017) [33]. Table I presents the asset classes considered for portfolio inclusion, and the functions of each asset. It can be observed that some of the portfolio's asset classes cover different stock markets to achieve global diversification. The chosen asset classes were the same as used by Wealthfront [32].

2) Macroeconomic Data: As can be seen in Table I, the portfolio's asset classes cover a variety of markets, not just the US equity market. Although the main focus is a US-based portfolio, the status of economies such as the Euro zone, the UK, Japan, and BRIC (Brazil, Russia, India and China) economies have a noticeable influence on the US market. Thus, the input dataset included as many economic indicators as possible for the US and other major economies in the developed and emerging markets, by including all of the economic indicators available on the Thomson Reuters Datastream [33].

3) Market Data: It is safe to assume that market variables generally influence asset returns. For this work, the LSTM neural network was exposed to market data consisting of major global equity market indices, commodities, major currencies' exchange rates, and the VIX index. The VIX index, also known as the market's "fear gauge", reflects the implied volatility of the S&P500 index options in the Chicago Board Options Exchange. It is a forward-looking measure of market expectations over the coming month's volatility [34]. This data should help to adjust the assets' weights during extreme or unexpected economic events over the analyzed time frame.

VI. Experiment

Classical passive portfolio management approaches CAPM, 3FM, EQWT (equally weighted portfolio rebalancing) were evaluated for this work in comparison with an LSTM-based approach. For CAPM and 3FM, the only input was the historical data of the various assets. For EQWT, the allocation in each of the 11 assets was set to 9.09%. This results in an equally weighted portfolio containing US stocks, foreign stocks, emerging market stocks, dividend stocks, real estate, natural resources, Treasury Inflation-Protected Securities (TIPS), municipal bonds, corporate bonds, emerging market bonds, and US bonds. CAPM assumes that all investors have all information at the same time, cannot influence prices, and can trade without paying for transactions [10]. 3FM assumes that value beats growth and that smaller companies do better than larger ones. Over the long term, investors must be able to handle extra short-term volatility for better long-term performance [11]. An initial version of the LSTM approach (LSTM in Table II and Figures 4 and 5), and a refined version developed through hyperparameter exploration (LSTM2 in Table II and Figures 4 and 5) are described in this work.

All computations were based on a fixed 24-month rolling window. A minimum of 5% (except TIPS have 0%) and maximum of 35% weight boundaries were used to achieve diversification and mitigate the MVO sensitivity problem. The use of multi-period optimization for CAPM and 3FM, and the rolling window method in MVO, enabled forward-looking estimations, and generated time-varying optimal weights for the portfolio assets.

Each model (CAPM, 3FM, EQWT, LSTM1, LSTM2) performed a rebalancing step on a monthly basis from January 2010 to August 2017, to maintain the target asset allocation scheme as determined by the investor's risk tolerance. The initial portfolio value for each approach was set to 100 USD, and recalculated monthly according to the consequences of the recommendations from the model for each approach. This traditional back testing methodology enables simple and straightforward comparisons of absolute returns. However, other statistical properties such as the Sharpe Ratio were observed to capture the risk adjusted return. Recall that a higher Sharpe Ratio generally indicates a more attractive investment.

The LSTM network from [5] (LSTM1) used the Mean Absolute Error (MAE) loss function. It used PCA to reduce the dimensionality of the market and macroeconomic data to 70 dimensions. It had a layer width of 16 and one hidden layer.

To develop the LSTM2 model reported in this work, the research team experimented with both MAE and Mean Squared Error (MSE) to observe which loss function would generate increased risk-adjusted returns. A variety of other parameters were experimented with as well. The number of dimensions outputted by the PCA function ranged from 10 to 200. The batch sizes used in the model were further explored, with values ranging from 1 to 512. Other values modified were the size of each epoch, the dropout rate, the width of the model, and the number of layers. Finally, early stopping was implemented to see if it would have a significant impact.

TABLE I. 11 assets considered for portfolio inclusion, and a description of their key benefits. The 10 year Sharpe ratio for each asset was obtained from Yahoo Finance for the period covering 2008 to 2018.

Assets Class	Ticker	Investment	Functions	Sharpe Ratio
US Stocks	IVV	iShares S&P500 Index ETF	Investment growth, long-term inflation protection, and tax efficiency	0.68
Developed Market Stocks	EFA	iShares MSCI EAFE Index ETF	Investment growth, long-term inflation protection, and tax efficiency	0.26
Emerging Market Stocks	EEM	iShares MSCI Emerging Market Index ETF	Investment growth, long-term inflation protection, and tax efficiency	0.24
Dividend Growth Stocks	VIG	Vanguard Dividend Appreciation Index Fund ETF	Investment growth, long-term inflation protection, and tax efficiency	0.74
US Government Bonds	SHY	iShares 1-3 year US treasury Bond ETF	Income, low historical volatility, diversification	0.75
Corporate Bonds	LQD	iShares iBoxx $ Inv Grade Corporate Bond ETF	Income, low historical volatility, diversification	0.74
Emerging Market Bonds	EMB	iShares JPMorgan USD Emerging Markets Bond ETF	Income, diversification	0.69
Municipal Bonds	MUB	iShares National Muni Bond ETF	Income, low historical volatility, diversification, tax efficiency	0.71
TIPS	TIP	iShares TIPS Bond ETF	Income, low historical volatility, diversification, inflation protection	0.45
Real Estate	VNQ	Vanguard REIT Index Fund ETF	Income, diversification, inflation protection	0.40
Natural Resources	XLE	Energy Select Sector SPDR Fund	Diversification, inflation protection, tax efficiency	0.22

The result of the described hyperparameter search was a model that used PCA to reduce market and macroeconomic data to 150 dimensions. The LSTM network had 3 hidden layers with a width of 128, and made use of early stopping.

LSTM1 and LSTM2 were both trained iteratively over a 92 month period, predicting the assets' future returns for the next month. These predictions were then passed along to the MVO calculations to reshuffle the asset weights. The training period increased by one month with each prediction. Each model was fitted using the Adam optimizer. This simulation was run ten times for each network.

VII. Performance Evaluation

The performance of classical passive portfolio management approaches (EQWT, CAPM, 3FM) and the active AI-based approaches described in this work (LSTM and LSTM2) were measured in terms of statistical properties (Table II) and percentage returns over time (Figure 4 and Figure 5). As can be seen, the results significantly outperformed both the classical methods and the previous LSTM model. Specifically, the observed mean annualized return for LSTM2 was 10.07%, whereas the other approaches generated less than half: 4.67% (CAPM), 3.49% (3FM), and 4.57% (EQWT), and 4.18% (LSTM). This much higher return was generated for a much improved risk-adjusted return. The mean Sharpe ratios of CAPM (0.46), 3FM (0.20), EQWT (0.54), and LSTM (0.43) were significantly lower that that of LSTM2 (0.97).

The results obtained for the conventional methods were in line with expectations. Whereas the return of individual assets under consideration may have been high (e.g. IVV returned over 100% growth), the risk of holding only that asset is unattractive in terms of the Sharpe ratio. Not knowing the future, any successful strategy should create risk-adjusted returns above the risk-free rate (e.g. fixed income treasury bonds) in order to justify the risk involved in following the strategy. LSTM2 was able to deliver high risk adjusted returns by holding growing assets with low volatility. The model was not explicitly trained to avoid volatility. Rather, it is likely that the training data led the model to converge upon a solution that allocates assets assessed to provide high return with high confidence into the portfolio. It is probable that the LSTM2 model has keyed in on low volatility by maximizing the probability of the expected return for each asset. In other words, the backpropagation “push” on the LSTM2 model to make high confidence asset allocations naturally discounts high volatility assets, adjusting their expected returns downward due to the relatively lower confidence.

Figure 4 shows that the return generated by LSTM2 was higher than the other approaches for all of the computed performance metrics. A more in-depth look is provided in Figure 5, where the growth of the portfolio returns for each simulation is tracked. The assets considered for the portfolio were the same for all models. The implication of a higher Sharpe ratio is that the volatility of the underlying assets held by LSTM2 was low. Another observation is that the returns for LSTM2 covered a broader range of outcomes than the single layer LSTM1. The higher variance in outcomes is tied in part to the fact that each month in the simulation requires the retraining of a new LSTM model, which trains using a new random seed in the random number generator. The large differences in the results of LSTM2 imply that chance does play a role in the future returns predictions. The future returns estimates approximate an unknown result with some error. This error aggregates each month resulting in different holdings for the models with different random seeds.

Figure 6 shows the average percentage returns of LSTM1 and LSTM2 along with the classical models, and the percentage returns of each individual asset in the portfolio. As can be seen, LSTM2 makes the best use of the more profitable assets, while still keeping a much lower Sharpe ratio than investing solely in those stocks. Figure 7 shows how the various Sharpe ratios did over time; each point is the rolling Sharpe ratio measured over the past year. LSTM2 consistently has a higher Sharpe ratio than the other methods.

It can be helpful to see how the different methods fared over the course of various economic events, to compare the performance and risk associated with each strategy. Figure 8 shows the percentage returns of the various methods over the course of the US downgrade in August of 2011 [35]. LSTM2 had the strongest recovery after the event, and as can be seen in Figure 9, and also maintained the highest rolling Sharpe ratio during this period of increasing market risk. Because there is a 5% minimum holdings requirement in the MVO for LSTM2, it is expected that the Sharpe ratio will broadly track market trends. Note in Figure 9, and more generally in Figure 7, that LSTM2 tracked the overall trend in the market, while reducing the risk exposure over other approaches for most of the simulated backtest period. During the time period of the US downgrade event, the LSTM2 model balanced the portfolio to favor US stocks, emerging markets, corporate bonds, real estate, and natural resources (IVV, EEM, LQD, VNQ, XLE). This balance of assets recovered slightly better during the recovery than other approaches. In addition, the LSTM2 model generated more consistent risk-adjusted returns than the other approaches.

LSTM2 also fared well during the European debt crisis of June 2012 [36]. It had a similar return on investment to the other methods (Figure 10), but was the only method to maintain a positive rolling Sharpe ratio throughout (Figure 11). During the time period of the European debt crisis, the LSTM2 model balanced the portfolio to favor stocks (developed market, emerging market, and dividend growth), corporate bonds, and real estate (EFA, EEM, VIG, LQD, VNQ). This balance of assets created a similar return as compared to the other approaches. Post crisis, the model shifted more into real estate, signalling a recovery in housing that was perhaps unrelated to the debt crisis in Europe.

During the Taper Tantrums of May 2013 [37], CAPM actually outperformed both LSTM models, both in percentage return and in rolling Sharpe ratio (Figures 12 and 13). This demonstrates that sometimes the LSTM2 model misjudges market movements and suffers in terms of risk exposure and returns as a result. This is the expected observation, as many of these crises are unforeseeable, and took the broader market by surprise. The Taper tantrums were a good example of market distortions driven by regulator action, rather than market fundamentals. This type of intervention should not be easily predicted by LSTM2, and indeed it underperformed during this period, generating a small positive return with a lower Sharpe ratio than CAPM and 3FM. During the time period of the Taper tantrums, the LSTM2 model balanced the portfolio to favor stocks (US, emerging markets, and dividend growth), TIPs, and real estate (IVV, EEM, VIG, TIP, VNQ). This balance of assets had worse returns than CAPM, though it still outperformed both the 3FM and LSTM1 methods. The LSTM2 model consistently had a lower Sharpe ratio than the classic approaches over those months. During this period, the model seems to have predicted a rise in inflation, and hedged by getting the portfolio into TIPs. This turned out to be wrong, but gives some interesting insight into what the model was expecting.

In Figure 14, LSTM2 can be seen to outperform the other models during the Russian financial crisis of December 2014 [38]. During the time period of the Russian financial crisis, the LSTM2 model balanced the portfolio to favor stocks (US, emerging markets, dividend growth), municipal bonds and real estate (IVV, EEM, VIG, MUB, VNQ). This balance of assets away from EFA (developed markets including European countries with Russian exposure) resulted in a much faster recovery after the event than the classic models. It also generated a significantly higher rolling Sharpe ratio (Figure 15).

Over the course of the China market crash of August 2015 [39], LSTM2 generated a negative return, and had a negative rolling Sharpe ratio, but still outperformed the other models (Figures 16 and 17). During the time period of the China market crash, the LSTM2 model balanced the portfolio to favor US stocks, dividend growth stocks, municipal bonds, real estate and natural resources (IVV, VIG, MUB, VNQ, XLE). The model was invested in MUB (municipal bonds) prior to the crash, and sold out to favor IVV and VNQ (US stocks and real estate) during and after the crisis. This balance of assets resulted in higher returns than other approaches, and fewer losses during the crisis. In addition, the LSTM2 model generated more consistent risk-adjusted returns than the other approaches as measured by the Sharpe ratio. The model seems to have predicted that the bond market yields would shrink due to government intervention, implying a shift in assets away from bonds and into stocks and other assets. This idea is further supported by the continuous drop in consistent returns among all of the observed strategy as shown in Figure 17.

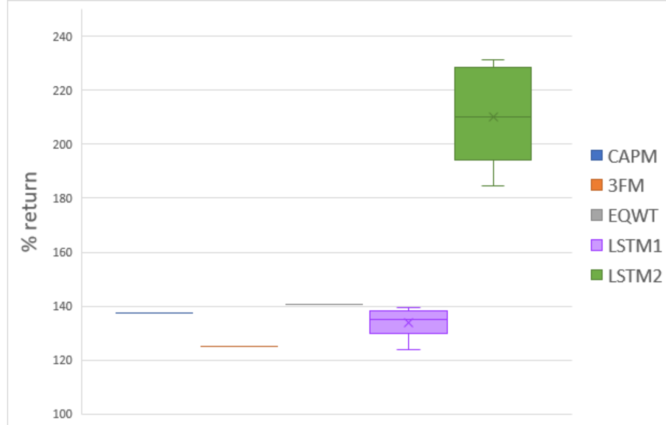


Figure 4. A comparison of the percentage return for each portfolio management method over 10 simulations

Finally, when Brexit happened in June of 2016 [40], the 3FM model had better percentage returns, but the LSTM2 model maintained a higher rolling Sharpe ratio over the same period (Figures 18 and 19). During the time period of Brexit, the LSTM2 model balanced the portfolio to favor US stocks, developed market stocks, emerging market stocks, real estate and natural resources (IVV, EFA, EEM, VNQ, XLE). There is significant UK exposure in EFA, and the model allocated away from EFA in the month following the crisis. In effect, the model missed the risk, as one would expect for such a low probability event. The asset allocation for LSTM2 resulted in lower returns than the 3FM method, but the LSTM2 model generated more consistent risk-adjusted returns than the other approaches as measured by the rolling Sharpe ratio.

Most returns generated by LSTM2 were generated during times of relative stability (non-crisis periods), as one would expect from a predictive model trained on historical data.

VIII. Conclusion

This work presented a neural network based approach to adaptive asset allocation recommendations. It combined LSTM and MVO models with data on historical prices, macroeconomic data, and market indicators. These signals were processed (e.g. scaling, PCA, rolling window) and then applied to address the portfolio management problem. The trained model was able to estimate the returns for a selected group

TABLE II. Statistics for the classical passive portfolio management approaches (EQWT, CAPM, 3FM) and active AI-based approaches (LSTM and LSTM2). The back testing period was from January 2010 to August 2017.

Statistics (Annualized)	CAPM	3FM	EQWT	LSTM	LSTM2
Mean Return	4.67%	3.49%	4.57%	4.18%	**10.07%**
Std Dev.	9.69%	10.64%	8.08%	9.26%	**10.13%**
R-f Average	0.24%	0.24%	0.24%	0.24%	**0.24%**
Sharpe Ratio	0.46	0.20	0.54	0.43	**0.98**

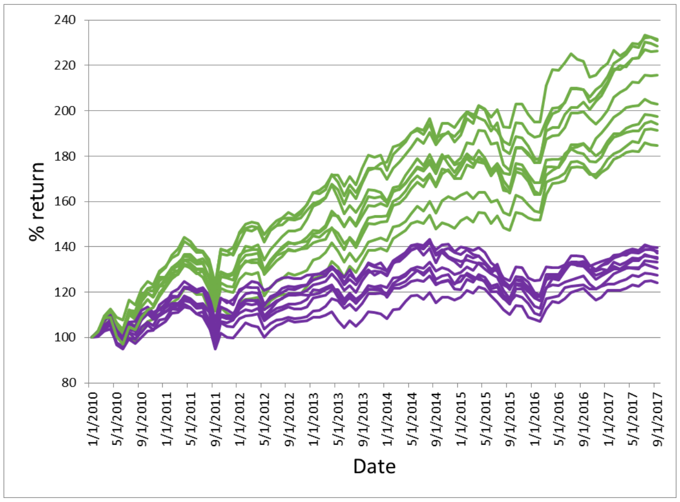


Figure 5. Percentage return comparison between the 10 simulations for LSTM1 (purple) and LSTM2 (green)

of diversified assets, and make investment recommendations with higher returns than traditional portfolio management approaches. The described system can assist in automating portfolio management. The asset allocation action recommendations were personalized to the portfolio under consideration, and were examined empirically in this work in comparison to standard portfolio management techniques. This work provides an empirical conclusion that an LSTM can provide better risk-adjusted returns than conventional strategic passive portfolio management approaches.

REFERENCES

[1] D. Shapiro, H. Qassoud, M. Lemay, and M. Bolic, "Unsupervised deep learning recommender system for personal computer users," in INTELLI 2017 : The Sixth International Conference on Intelligent Systems and Applications, 2017.

[2] J. Schlesinger, "How chatbots and artificial intelligence are changing wealth management," May 2017. [Online]. Available: https://www.theglobeandmail.com/report-on-business/will-computer-deep-learning-lead-to-higher-earnings/article34842587/

[3] "Horizons ETFs launches canada's first ETF driven by A.I." Nov 2017. [Online]. Available: https://www.newswire.ca/news-releases/horizons-etfs-launches-canadas-first-etf-driven-by-ai-654424823.html

[4] C. Metz, "The rise of the artificially intelligent hedge fund," Jun 2017. [Online]. Available: https://www.wired.com/2016/01/the-rise-of-the-artificially-intelligent-hedge-fund/

[5] S. Obeidat, "Adaptive asset allocation optimization using artificial intelligence," Master's thesis, Royal Roads University, 2017.

[6] F. Cai, N.-A. LeKhac, and M.-T. Kechadi, "An integrated model for financial data mining," in Multi-disciplinary Trends in Artificial Intelligence, C. Sombattheera, N. K. Loi, R. Wankar, and T. Quan, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 306–317.

[7] B. G. Malkiel, A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing. WW Norton & Company, 2011.

[8] F. J. Fabozzi, "The legacy of modern portfolio theory," CFA Digest, vol. 33, no. 2, 2003, pp. 50–52. [Online]. Available: https://doi.org/10.2469/dig.v33.n2.1273

[9] H. Markowitz, "Portfolio selection," The journal of finance, vol. 7, no. 1, 1952, pp. 77–91.

[10] W. F. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk," The journal of finance, vol. 19, no. 3, 1964, pp. 425–442.

[11] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," Journal of financial economics, vol. 33, no. 1, 1993, pp. 3–56.

[12] W. Commons, "File:markowitz frontier.jpg — wikimedia commons, the free media repository," 2015, [Online; accessed 30-January-2018]. [Online]. Available: https://commons.wikimedia.org/w/index.php?title=File:Markowitz_frontier.jpg&oldid=164635807

[13] C. L. Dunis, P. W. Middleton, A. Karathanasopolous, and K. Theofilatos, Artificial Intelligence in Financial Markets: Cutting Edge Applications for Risk Management, Portfolio Optimization and Economics. Springer, 2016.

[14] T. Marwala, Economic modeling using artificial intelligence methods. Springer, 2013.

[15] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural networks, vol. 2, no. 5, 1989, pp. 359–366.

[16] P. D. McNelis, Neural networks in finance: gaining predictive edge in the market. Academic Press, 2005.

[17] L. Medsker, "Design and development of hybrid neural network and expert systems," in Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on, vol. 3. IEEE, 1994, pp. 1470–1474.

[18] R. Vince, The Handbook of Portfolio Mathematics: Formulas for optimal allocation & leverage. John Wiley & Sons, 2008, vol. 400.

[19] J. L. Treynor, "How to rate management of investment funds," Harvard business review, vol. 43, no. 1, 1965, pp. 63–75.

[20] M. C. Jensen, "Risk, the pricing of capital assets, and the evaluation of investment portfolios," The Journal of Business, vol. 42, no. 2, 1969, pp. 167–247.

[21] F. D. De Freitas and A. R. De Almeida, "Portfolio selection with predicted returns using neural networks," in IASTED International Conference on Artificial Intelligence and Applications, 2001, pp. 99–103.

[22] G. Jang and F. Lai, Intelligent trading of an emerging market. John Wiley & Sons Inc, 1994.

[23] P.-C. Ko and P.-C. Lin, "Resource allocation neural network in portfolio selection," Expert Systems with Applications, vol. 35, no. 1, 2008, pp. 330–337.

[24] R. Raei, "Risky portfolio selection through neural networks," The Iranian Accounting and Auditing Review, vol. 13, no. 46, 2007, pp. 71–85.

[25] F. Black and R. Litterman, "Global portfolio optimization," Financial analysts journal, vol. 48, no. 5, 1992, pp. 28–43.

[26] P. Jorion, "Portfolio optimization in practice," Financial Analysts Journal, vol. 48, no. 1, 1992, pp. 68–74.

[27] D. Davidian, "Feed-forward neural network," Aug. 1 1995, US Patent 5,438,646.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, 1997, pp. 1735–1780.

[29] S. Il Lee and S. J. Yoo, "A new method for portfolio construction using a deep predictive model," in Proceedings of the 7th International Conference on Emerging Databases, W. Lee, W. Choi, S. Jung, and M. Song, Eds. Singapore: Springer Singapore, 2018, pp. 260–266.

[30] L. I. Smith et al., "A tutorial on principal components analysis," Cornell University, USA, vol. 51, no. 52, 2002, pp. 12–20.

[31] Q. C. Chu and D. N. Pittman, Treasury Inflation-Protected Securities. Boston, MA: Springer US, 2013, pp. 257–262. [Online]. Available: https://doi.org/10.1007/978-1-4614-5360-4_8

[32] "Wealthfront investment methodology white paper." [Online]. Available: https://research.wealthfront.com/whitepapers/investment-methodology/

[33] "Thomson reuters datastream." [Online]. Available: https://financial.thomsonreuters.com/en/products/tools-applications/trading-investment-tools/datastream-macroeconomic-analysis.html

[34] A. Schrager, "The stock market's "fear gauge" is at a 23-year low, and traders are terrified," May 2017. [Online]. Available: https://qz.com/979370/volatility-and-the-vix-index-the-stock-markets-fear-gauge-is-at-a-23-year-low-and-traders-are-terrified/

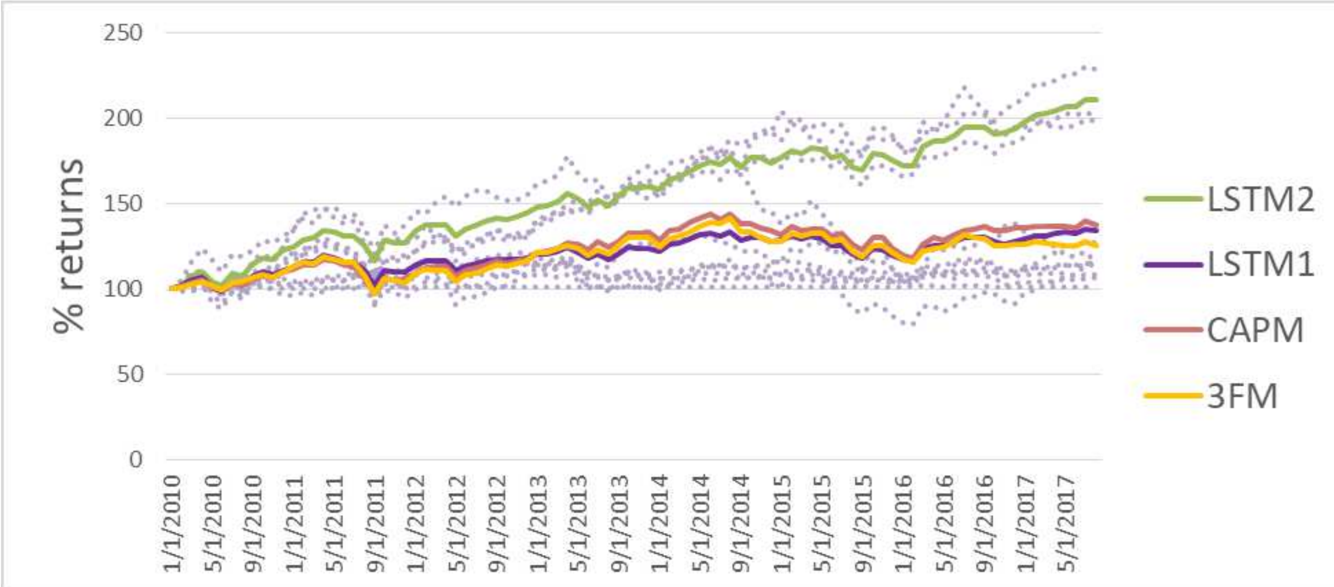


Figure 6. The % return for each underlying asset during the testing period was reported with purple dots. The average % returns for LSTM1 and LSTM2 were overlaid onto the graph. The % returns of CAPM and 3FM were also overlaid onto the graph.

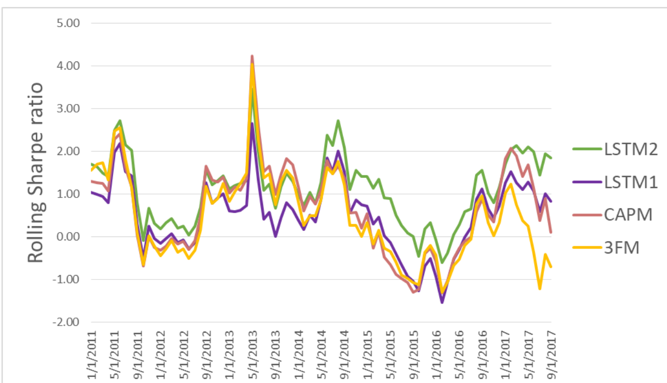


Figure 7. Average rolling Sharpe ratios for LSTM1 and LSTM2 along with rolling Sharpe ratios for CAPM and 3FM

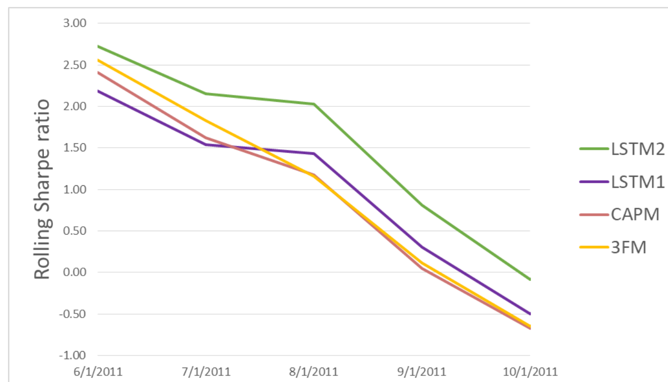


Figure 9. Rolling Sharpe ratios during the timeframe of the US downgrade of August 2011

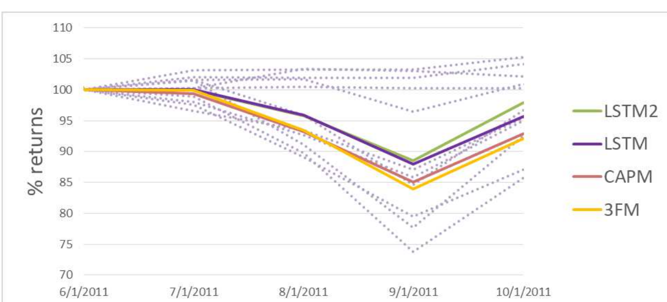


Figure 8. Percentage returns during the timeframe of the US downgrade of August 2011. The % return for each underlying asset during the testing period was reported with purple dots.

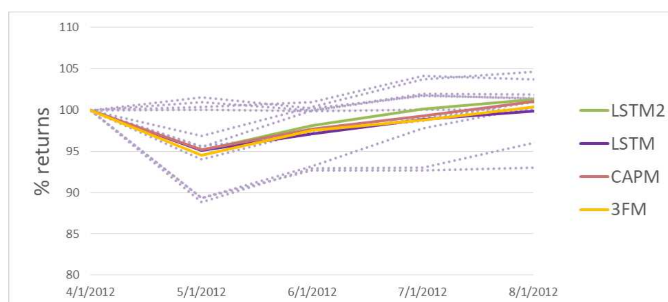


Figure 10. Percentage returns during the timeframe of the European debt crisis of June 2012. The % return for each underlying asset during the testing period was reported with purple dots.

[35] C. Riley, "S & P downgrades U.S. credit rating," Aug 2011. [Online]. Available: http://money.cnn.com/2011/08/05/news/economy/downgrade_rumors/index.htm

[36] S. Steiner, "Timeline of European debt crisis," Oct 2012. [Online]. Available: http://www.bankrate.com/banking/timeline-of-european-debt-crisis/

[37] A. Kaletsky, "The markets and Bernankes "taper tantrums"," Nov 2013. [Online]. Available: http://blogs.reuters.com/anatole-kaletsky/2013/09/

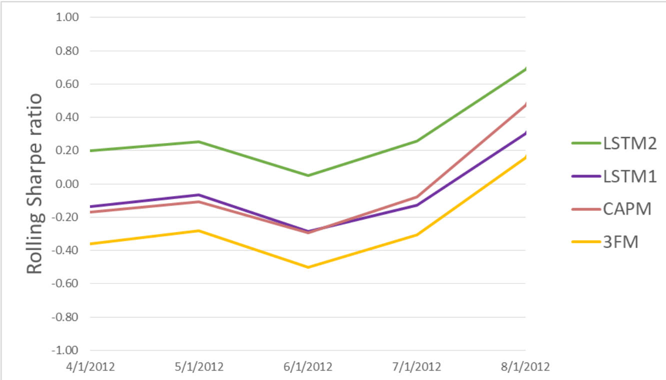


Figure 11. Rolling Sharpe ratios during the timeframe of the European debt crisis of June 2012

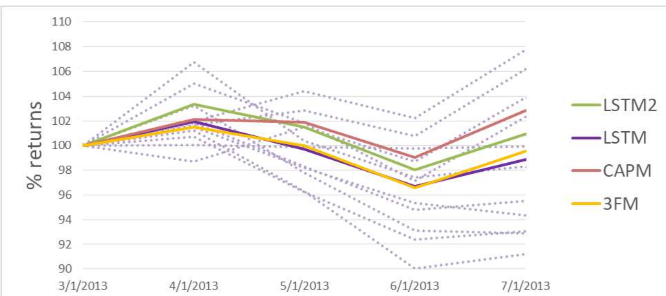


Figure 12. Percentage returns during the timeframe of the Taper tantrums of May 2013. The % return for each underlying asset during the testing period was reported with purple dots.

19/the-markets-and-bernankes-taper-tantrums/

[38] N. Kitroeff and J. Weisenthal, “Here’s Why the Russian Ruble Is Collapsing,” Dec 2014. [Online]. Available: https://www.bloomberg.com/news/articles/2014-12-16/no-caviar-is-not-getting-cheaper-everything-you-need-to-know-about-the-russian-ruble-collapse

[39] J. Mchugh, “China ‘Black Monday’ Timeline: The Chinese Stock Market Crash And How It Happened,” Aug 2015. [Online]. Available: http://www.ibtimes.com/china-black-monday-timeline-chinese-stock-market-crash-how-it-happened-2065173

[40] G. Rachman, “Brexit and the making of a global crisis,” Jun 2016. [Online]. Available: https://www.ft.com/content/5490d754-3a0e-11e6-9a05-82a9b15a8ee7

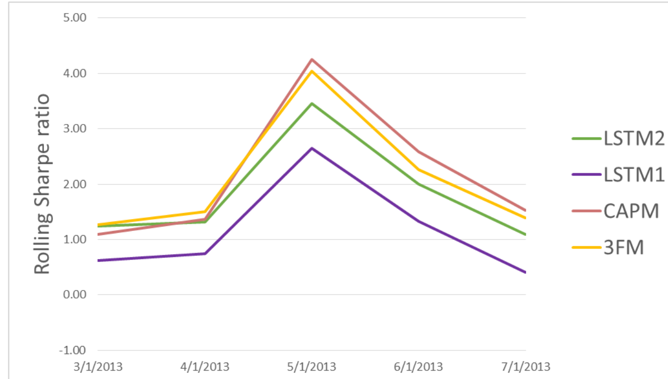


Figure 13. Rolling Sharpe ratios during the timeframe of the Taper tantrums of May 2013

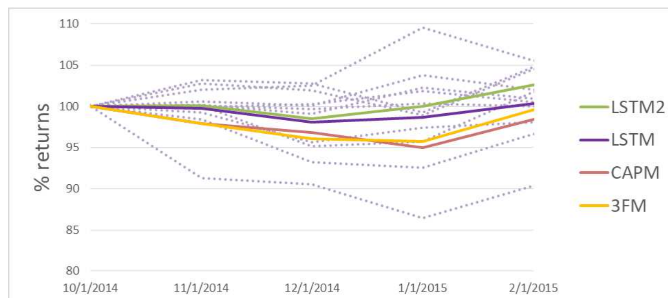


Figure 14. Percentage returns during the timeframe of the Russian debt crisis of December 2014. The % return for each underlying asset during the testing period was reported with purple dots.

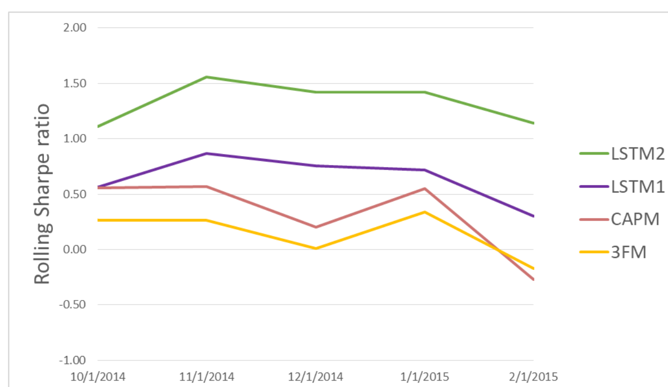


Figure 15. Rolling Sharpe ratios during the timeframe of the Russian debt crisis of December 2014

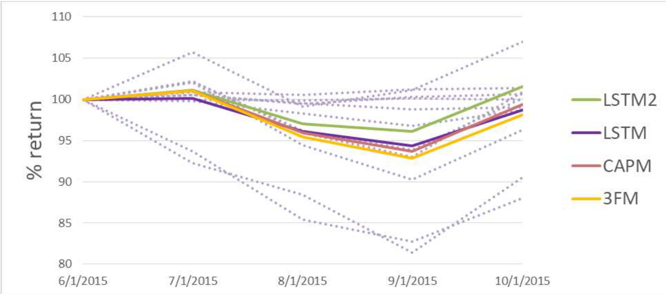


Figure 16. Percentage returns during the timeframe of the China market crash of August 2015. The % return for each underlying asset during the testing period was reported with purple dots.

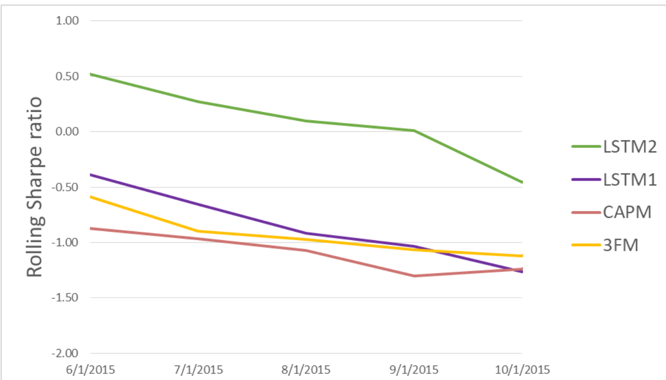


Figure 17. Rolling Sharpe ratios during the timeframe of the China market crash of August 2015

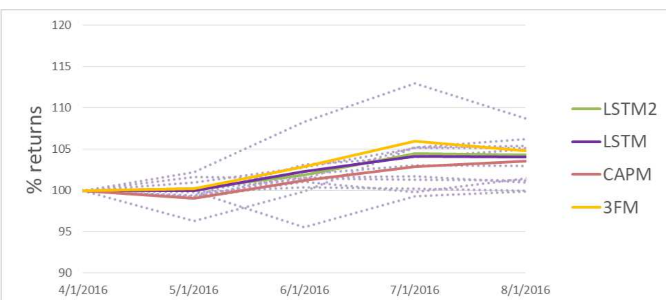


Figure 18. Percentage returns during the timeframe of Brexit, June 2016. The % return for each underlying asset during the testing period was reported with purple dots.

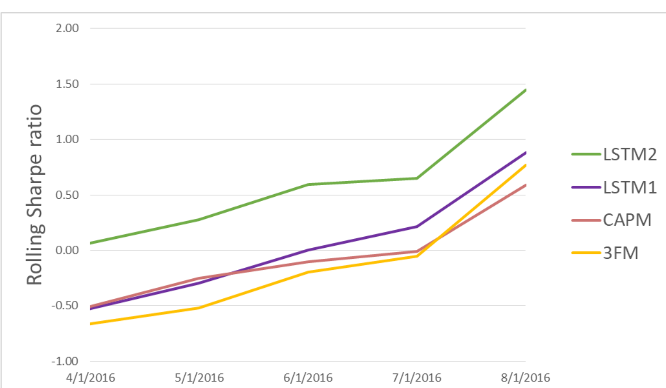


Figure 19. Rolling Sharpe ratios during the timeframe of Brexit, June 2016

Cartesian Systemic Pulsation –

A Model for Evolutive Improvement of Incomplete Symbiotic Recursive Systems

Marta Franova, Yves Kodratoff
LRI, UMR8623 du CNRS & INRIA Saclay
Bât. 660, Orsay, France
e-mail: mf@lri.fr, yvekod@gmail.com

Abstract— **Developing and using Symbiotic Recursive Systems (SRS) concerns various incompletely defined domains requiring handling prevention and control. This means that, in the innovation process, it is important to avoid decline and obsolescence as it happens with modern paradigms of innovation. This paper presents the specificities of SRS that call for a new model of evolutive improvement. Such a new model needs to handle prevention and control in SRS as well as in the improvement process. The paper presents such a model we call 'Pulsation'. Symbiotic Recursive Pulsative Systems are then SRS including the process of Pulsation.**

Keywords-pulsation; Symbiotic Recursive Pulsative Systems; systemic recursion; Ackermann's function; security; practical completeness.

I. INTRODUCTION

We work under the hypothesis that human knowledge is and hopefully will always be incomplete. This incompleteness favours innovation and discovery. One distinctive feature of modern innovation is acceptance of a rapid obsolescence and decline of the technologies or systems it develops. In many cases, however, this last feature cannot be accepted. As an example, dealing with symbiotic recursive incomplete systems needs a robustness which asks for special attention. This paper addresses this problem in presenting a new model for an evolutive development of new symbiotic recursive systems or technologies. It has been first introduced in [1] where we called it: Pulsation. Its basic features are – at each step of development –

- a kind of timelessness of previous achievements (no obsolescence);
- possibility for future new improvements (no decline);
- focus on prevention and control (particular rigor).

It is true that modern science and the philosophy of innovation tend to deal with

- synergy and modularity instead of symbiosis,
- non-recursive complexities instead of recursion, and
- constant change or mutations instead of what we call Pulsation.

While these modern science notions are extremely useful and relevant for many real-world applications, they however cannot replace symbiosis, recursion and Pulsation without harmful consequences. This means that the approaches of these two groups of notions are complementary and non-competitive.

For convenience, we shall call *first group* the one including the notions

- symbiosis
- recursion
- Pulsation

and *second group* that of including the notions

- synergy, modularity
- non-recursive complexities
- change, mutations.

This paper gives below a systemic description of the first group of notions.

In order to illustrate Pulsation in action, we present it in the framework of Symbiotic Recursive Systems (SRS). We show that these systems are particularly suited to represent potentially incomplete SRS that formalize real-world applications. An example of such a real-world application will be given.

We shall also present reasons for naming *Cartesian Intuitionism* a systemic paradigm based on the first group and *Newtonian approach* a systemic paradigm based on the second group of notions. We have called *Symbiotic Recursive Pulsative Systems* (SRPS) the systemic nucleus that is the basis of Cartesian Intuitionism. One of our goals is to introduce this notion that is, to the best of our knowledge, not referred at elsewhere.

The paper is organized as follows.

Section II specifies the notion of symbiosis as understood in this paper. Since symbiosis is related to our understanding of the notion of theory we shall introduce, with help of symbiosis, a difference between formal and deductive theory. Section III introduces recursion as a way of representing action, control and prevention. It explains what we mean by systemic recursion. Section IV introduces the notion of oscillation as representation of one-level creation process used in Pulsation presented in Section V. Section VI presents a motivation for introducing the systemic difference between Newtonian approach and Cartesian Intuitionism. It will become clear why the latter expression is used to describe the systemic science relevant to SRPS. Section VII presents an application for which Pulsation is relevant. Section VIII relates ancient systemic thinking to Pulsation and SRPS.

II. SYMBIOSIS : 'VITAL INTERDEPENDENCE'

Symbiosis is a particular composition. In this section we shall give a definition of symbiosis as used in this paper. We shall also compare symbiosis to other kinds of composition, namely synergy and fusion.

By symbiosis we understand a separation-sensitive composition of two or several parts. This means that a separation of one or several parts leads to extinction or irrecoverable mutilation of the whole and all the involved parts, as will be illustrated below.

In contrast to this, by synergy we understand a composition of parts that is not separation-sensitive. Sometimes, synergy is called also modular composition.

In case of fusion, the resulting composition is homogenous. It does not allow to recognize the involved parts, they are blended together, such as in fusion of metals.

We shall now give several examples.

A. *Pictorial Symbiosis*

Let us consider the following version 'two-women-in-one' of 'Devinettes d'Épinal' (see also [18]):

Figure 1. 'Two-women-in-one' picture.

Several overlapping features may reveal two different human faces. In other words, this picture is a composition of two parts. The important point is that the features necessary to see a 'young' or an 'old' face are common to both visions (i.e., parts), though they may be differently interpreted. Here, the feature 'little chin' in one is interpreted as 'big nose' in the other, the ear of one is interpreted as the eye of the other, the necklace of one becomes the mouth of the other, and the couple one eyelash + small nose of one becomes the two eyelashes of the other. If we withdraw from the picture all of these common features, as shown below, then a human face recognition 'system' may still reconstruct a human face though it becomes unable to rebuilt these common features of different interpretations, the 'two-faces-in-one' picture is thus destroyed and irrecoverable. This represents a symbiotic pictorial occurrence of both faces, that is, there exist a subset of features of these two parts (here, four of these features, but this is not necessary) such that deleting them from one occurrence induces an unrecoverable loss of the picture intent.

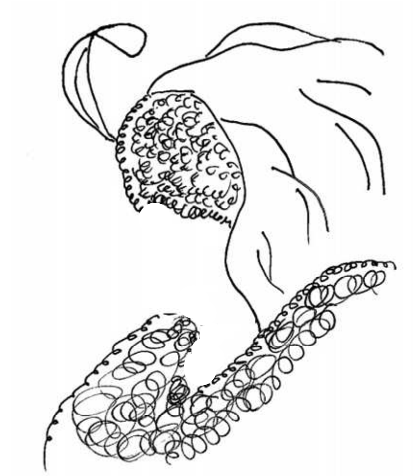


Figure 2. Mutilated 'two-women-in-one' picture.

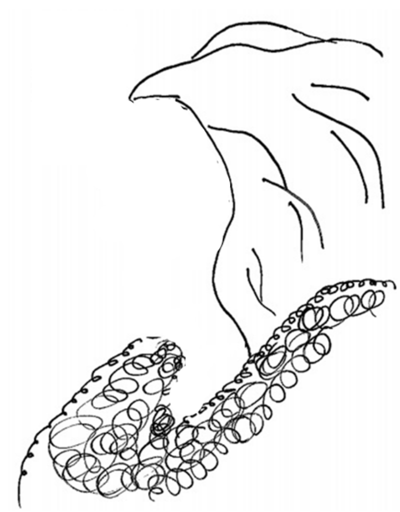


Figure 3. Irreversible mutilation of 'two-women-in-one' picture.

The feature 'hair decorated with a feather', as in Figure 2, is common to the two faces, and may evoke an incomplete female face. If we withdraw this last feature, as in Figure 3, then a human recognition system is at lost at recognizing something really significant (a strange bird, perhaps?). While Figure 1 represents several overlapping features of two human faces, from the point of view of symbiosis, it is important to understand it as a composition of two parts (old woman, young woman). Therefore, it is relevant that eliminating one part (and not one feature) leads to the destruction of the whole and the remaining parts. Here, for instance, Figure 3 can be seen as a result of eliminating old woman. Young woman disappears as well.

B. Representational Symbiosis

A careful study of primitive notions in Euclid's geometry [20] shows that these notions are symbiotic. This means that eliminating even one notion would either render meaningless the resulting system (i.e., extinction of the resulting system) or the meaning of the resulting system would be completely different (i.e., irrecoverable mutilation).

It should be noted here that the example of Euclid's geometry illustrates well the fact that the constituents of a symbiotic system need to be handled as symbiotic in the construction process of such a system. However, after its successful final creation, the use of these notions may, in some cases, be modular. For instance, when we use the notion of point while working in Euclid's geometry, we do not need to be aware of the symbiotic dependence of this notion on other notions of the same geometry. However, such awareness was necessary for Euclid when he created this geometry. This point is further illustrated in Section II.D.

C. Intentional symbiosis

As far as intentional symbiosis is concerned, we consider it exclusively in relation to human creations. A detection of intentional processes in nature is out of the scope of this paper.

We could perceive a slight glimpse of intentional symbiosis in all the above, somewhat static, examples. Even though present, a rather procedural character of intentional symbiosis was not mentioned. In Section III., we shall give an example of the construction of Ackermann's function, where such an intention can be easily described. As far as real-world applications are concerned, usually their development is driven by synergic thinking. This synergic thinking is convenient when all the tools are available and the resulting system consists in a novel composition of these tools. When it happens that new basic tools need to be invented, symbiotic thinking opens the way to new conceptual switches enabling some breakthrough from the usual thinking. This underlines why symbiotic thinking is an asset for creating new technologies.

D. Deductive and Formal systems

The above example of representational symbiosis, namely that of Euclid's geometry, inspires us to introduce a difference between a deductive and a formal system. Indeed, when a formal system is considered in science, its consistence is considered in terms of non existence, in this system, of a proof for a formula as well as for its negation. By deductive system we understand a system developed with a concrete real-world application as a model. This means that the consistence of deductive systems is asserted by the existence of a concrete model. In fact, a deductive system is in our work viewed as a result of development of a relevant axiomatic system for a particular intended application. In the final stage of development, a deductive system can be viewed as a formal system, however, its completeness or incompleteness is not viewed from a theoretical point of view but from the point of view of a pragmatic evaluation. For instance, Gödel has shown the theoretical incompleteness of the set 0, 1, 2, …. However, when we consider natural numbers NAT as a deductive theory the intended model of which are the numbers we all *use*, i.e., the numbers represented by Peano's arithmetic, we can consider NAT as being practically complete. Indeed, in this practical case, we need to consider symbiotic relationship of numbers and axioms defining the addition, the multiplication, and the induction principle. In other words, for deductive systems we introduce the notion of practical completeness. Practical completeness means that we all agree on the interpretation (i.e., the model) that is considered. Usually, this is allowed when there is no ambiguity as to the exact meaning of the notions in their practical manipulations. In order to illustrate the 'practical completeness' of natural numbers, think how all computer driven money exchanges in the world use the same intended model of the natural numbers. When, on the contrary, such an ambiguity is possible this indicates that the developed deductive system is incomplete. Selecting a concrete version of the intended model will only be possible when the corresponding notions have been (at least partially) completed through a relevant completion of the developed deductive theory. By partial completion of definitions of notions, we mean definitions that guarantee practical completeness of the considered system.

In order to illustrate the informal (or incomplete) character of notions in incomplete theories, let us recall that, in a geometry obtained from Euclid's geometry by eliminating the postulate of parallels, a triangle can be defined. However, in this incomplete 'theory', the sum of the triangle angles may differ from 180°. This means that the notion of triangle is incompletely defined in this particular purged (or mutilated) Euclid's geometry. In practice, it means that an informal definition covers several possible different interpretations of each 'defined' object. Thus, deductive theories are characteristic by their origin in a concrete application and their incompleteness is not a limitation, when practical completeness only is requested.

Our model of Pulsation presented in Section V is developed with the aim to guarantee a rigorous development, or completion, of deductive systems that corresponds to intended real-world technological applications. It will become clear in Section V that Pulsation deals also with another feature of practical completeness, namely the availability of solutions for practical problems that can be described in the theory.

The notion of symbiosis is in our work an emergent notion. This means that the specification of symbiosis presented in this section will have to be refined symbiotically with the future development of theory of SRPS. Namely, we still need to develop and present strategic and practical aspects of the creation of symbiotic systems. In [13] we call Cartesian Systemic Emergence this process. We illustrate there also one particular strategic feature on a simple example.

III. Recursion

In this section we present a minimal, but sufficient basis for understanding *systemic recursion* for Symbiotic Recursive Pulsative Systems (SRPS). Moreover, we will illustrate how recursion represents not only actions, but also particular forms of control and prevention.

Mathematical and computational recursion handle recursion from formal or programs efficiency points of view (see [15], [14]). Recursion in these cases is a known tool and not a science. In contrast to this, systemic recursion is a *science of know-how for creating recursive systems* that are useful for real-world applications in various domains. We shall point out the main features of systemic recursion through out this paper.

A. Preliminary definitions and notions

It is known that recursion is a particular way to represent, by a finite set of rules, potentially infinite systems and processes (actions and creations). These rules are expressed in terms of basic action or creation operators called constructors.

In Mathematics and Computer Science such constructors are usually known, available or easily attainable in standard know-how. In contrast to this, know-how of systemic recursion lies in a progressive invention of on-purpose constructors of a goal system in dependence with progressive invention of a *formal specification* of the goal system. Indeed, in systemic recursion, we start to build a system from an *informal specification* of the goal system. The notion of Pulsation presented in Section V is important for understanding the systemic emergence of a formal system from its informal specification.

By informal specification of a real-world application we mean that it is not yet a formalized description. It is, for instance, the case for technological visions or some pragmatic formulations of technological needs. An informal specification is usually not formulated by a mathematician or by a computer scientist. It is formulated by a visionary person or by experts expressing a need for some new solving tools in their domain. Informal specification describes a rather vague 'what' of the intended application, tool or system.

When an informal specification of a technological vision is known, the 'how' of its implementation is usually not available and may even be unknown or impossible in standard know-how. Therefore, in systemic recursion, we speak of creation or even emergence rather than of development. Since not all the constructors of the intended system are known at the beginning, a long period of preliminary research of a sufficient set of relevant constructors always precedes the implementation of an experimental prototype. The final development might even require some complementary inventions.

The goal of systemic recursion is to pass from an informal specification to a satisfactory formal specification of the goal together with the relevant on-purpose knowledge and know-how, i.e., the 'how' or 'procedural science' of the actual development. A formal specification expresses all the knowledge necessary for a final implementation of the 'what' that has been at first only informally specified.

This means that, in systemic recursion, a formal specification is an agreed upon compromise between the visionary and the developers of the considered informal specification. This compromise is built up progressively. It cannot be created in advance. This is because, in systemic recursion, research is pluridisciplinary in the sense that it crosses the traditional boundaries between disciplines and it progressively constructs its own on-purpose knowledge, know-how and boundaries. This on-purpose systemic knowledge is symbiotic.

In the following section we shall show how recursive actions may be considered as a way to represent not only actions but also a particular kind of control.

B. Representation of a particular control

In this section we shall present an example that illustrates how recursion captures in itself, by symbiotic dependency, all the secondary effects of a simple recursive procedure computation. Let us point out that we emphasize here symbiotic information, not symbiotic computation.

Let us consider the following simple problem. On a sufficiently big table consider a stack of blocks a, b, c, d and e as shown in Figure 4.

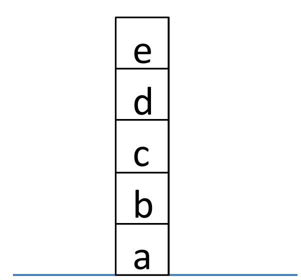


Figure 4. A stack of blocks.

We say that a block m is clear if there is no other block on m. (In Figure 4. block e is clear.) There can be at most one block on the top of the other. If n is on the top of m we say that n is top of m written as: n = top(m). Let us consider the following procedure makeclear:

```
makeclear(x) =
   if x is clear then end
   else
       if     top(x) is clear
       then   put(top(x)) on table
       else   first makeclear(top(x))
              and
              then put(top(x)) on table
```

It can easily be checked that makeclear(b) results not only in clearing the block b but also in the situation where blocks c, d and e are clear and on the table. This means that the procedure makeclear contains in its description not only its direct effects (such as: the block b is cleared) but also the full description of all the secondary effects of any action performed. In Figure 5 these secondary effects are that the blocks c, d and e are on the table.

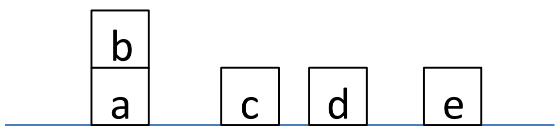


Figure 5. The environment after clearing b.

For some primitive recursive procedures the secondary effects do not modify the environment, but this should not be a barrier for a general perception of primitive recursive procedures as invisible procedural 'seeds' containing symbiotically related the effects (i.e., the results of the computations) and the secondary effects (i.e., the consequences of the computation of a particular value). Therefore, implementing recursive procedures is interesting in all the environments where the control over the secondary effects is important.

It may be not straightforward to see in what sense we speak of symbiotically related effects and secondary effects. Let us recall that symbiosis of information means here that if we take away one piece of information, the global information becomes distorted, or mutilated.

Let us consider therefore the instruction makeclear(b). This means that, step-wise, all blocks above b have to be put on the table leading to the result that the blocks c, d and e are on the table. Let us suppose that we take away the information that the block d is on table. However, in the progression of the procedure, if we call makeclear(b) first, we have to call makeclear(c), then makeclear(d). Since top(d) is e and it is clear, we may put e on the table. d = top(c) is now clear, but we cannot put it on the table since this information has, by our assumption, been withdrawn. This means that the whole process is stuck and we then cannot put block c on the table. So, the resulting information is mutilated since it does not express all the potential of the procedure makeclear. Moreover, there is no evidence that the information that the block e is on the table is related to the procedure makeclear. It could, in principle, be related to some other procedures.

The above procedure makeclear is an example of primitive recursion. A recursion that is not primitive goes even further in representing symbiotically information that concerns control, rigor and reproducibility. Ackermann's function is the simplest example of a non-primitive recursive function. Therefore, it is a suitable representative for explaining how non-primitive recursion modelizes a particular kind of Pulsation in SRPS.

In the following part we shall give a formalized presentation of the Pulsation starting by a presentation of a construction procedure that generates Ackermann's function. It will become clear how this construction and the notion of Pulsation are linked together.

C. A Construction of Ackermann's Function

The idea to modelize Pulsation by Ackermann's function comes from the understanding of how this function may be constructed. The practical use of this function becomes then exploitable by a 'simplification' of the computation of its values using the knowledge of its construction process.

Let 'ack' be Ackermann's function defined, as in [19], by its standard definition, i.e.,

$$ack(0,n) = n+1$$
$$ack(m+1,0) = ack(m,1)$$
$$ack(m+1,n+1) = ack(m,ack(m+1,n)).$$

We shall show here how this function can be constructed.

Since ack is a non-primitive recursive function, by definition of non-primitive recursion, it is a particular composition of an infinite sequence of primitive recursive functions. We shall thus define a function ack' as a particular composition of an infinite sequence of primitive recursive functions and it will become clear why the definitions for ack and for ack' are identical.

By definition, each primitive recursive function f is a composition of a finite number of primitive recursive functions and of f itself. Let us therefore construct such an infinite sequence of primitive recursive functions f_0, f_1, f_2, …, f_n, f_{n+1}, …. We define

$$f_0(n) = n+1$$
$$f_{i+1}(n+1) = f_i(f_{i+1}(n))$$

for each i from 0, 1, 2 …. We are thus able to define a new function ack' as follows: $ack'(0,n) = f_0(n)$ and $ack'(m+1,n+1) = f_{m+1}(n+1)$. Note that ack'(m+1,0) is not yet defined. Since we want ack' to be a non-primitive recursive function, we need to guarantee that it cannot be reduced to any of f_i. In order to do so we shall simply perform a diagonalization [15] on this infinite sequence of functions by defining

$$f_{i+1}(0) = f_i(1).$$

In other words, we define

$$ack'(m+1,0) = f_m(1).$$

By this construction, we see that f_{i+1} is more complex than f_i for each i. It is obvious that

$$ack'(m,n) = ack(m, n) = f_m(n).$$

This construction is at the same time a guarantee that ack is not primitive recursive, since it is indeed a composition of an infinite sequence of primitive recursive functions each of them more complex than those before it and ack cannot be reduced to any one of them. As a by-product, we have thus simplified also the standard presentation of the non-primitive character of ack which is usually done by a proof by a projection of Ackermann's function ack into a sequence of primitive recursive functions $a_m(n)$ = ack(m,n) and showing that ack grows more rapidly than any of these primitive recursive function (see [19]). The difference thus lies in our use of an indirect construction (instead of a projection) and relying on a progressive diagonalization. To our best knowledge, this construction with a use of progressive diagonalization was not presented so far. Note that the notion of Pulsation that refers to this construction of Ackermann's function has no relation to measures of the computation complexity of a function, such as Ritchie's hierarchy [17]. Complexity and efficiency are thus out of the scope of this paper.

D. Disentangling Ackermann's Function

The above construction of Ackermann's function shows immediately that the computation of its values, for given m and n, using non-primitive recursive definition can be 'simplified' - or, rather, replaced - by a definition of m primitive recursive functions obtained by a suitable macro-procedure.

Our recursive macro-procedure will simply compute, step by step, each of the values $f_{i+1}(0)$ in advance and will define the whole f_{i+1} with this already computed value. This may not lead to a fast computation but we are not concerned now with computational efficiency of this way of proceeding, only by its practical feasibility and reproducibility.

We define a macro-procedure, we call ack_macro. It uses the standard LISP procedures add_to_file and load_file. The procedure add_to_file(*text*,*F*) adds the *text* at the end of file *F*. The procedure load_file(*F*) loads file *F* in order to make computable the functions written in this file. We create at the start an auxiliary file F that stores the functions f_i generated by ack_macro. Our macro-procedure ack_macro(m,n) uses the infinite sequence of functions defined above as being representative of Ackermann's function.

```
Step 1:
   text:= { f0(n) = n+1 }
Step 2:
   Create file F (empty at start) and
           add_to_file(text,F)
   load_file(F)
Step 3:
   i:=0
   aux:= compute the value of fi(1)
Step 4:
   text := { fi+1(0)= aux
           and fi+1(n+1)= fi(fi+1(n))}
   add_to_file(text,F)
   load_file(F)
   aux := compute the value of fi+1(1)
   i:= i+1
   if i < m
   then Go to Step 4
   else stop
```

Figure 6. A macro-procedure for computing particular values of ack

ack_macro(m,n) is now completed and file F collects the definitions of m primitive recursive functions. We are now able to compute ack(m,n) = $f_m(n)$, where the definition of f_i from file F is used for all i = 0, 1, …, m.

On internet, we can find several programs that compute Ackermann's function values faster and much further than our presented macro-procedure. The problem with those programs is that they are based on the knowledge that Ackermann's function can be represented as a generalized exponentiation function. The advantage of our presentation lies in its suitability for practical purposes in the following sense. As we shall see with the notion of Pulsation, complex real-world problems may require modelization by non-primitive Ackermann's like programs that will not be reducible to an arithmetic generalized exponentiation. In other words, it is useful to consider non-primitive recursion that is not defined for natural numbers only but which is defined for all practically useful and exploitable recursive systems.

E. Prevention and Control in Recursion

We have seen above, in the example of program makeclear, that primitive recursion captures the effects (the value of the computation) and the secondary effects (the consequences of the computation that are in fact the intermediary values generated by the same procedure). We have also seen that the non-primitive recursive Ackermann's function is obtained using a diagonalization procedure. This diagonalization brings forward complementary information about the process of this symbiotic information in recursion. Since diagonalization is a meta-level procedure, we understand this complementary information as a kind of meta-level prevention from primitive recursion reducibility. While lack of control is accepting to ignore some secondary effects of the computation, lack of prevention is accepting to ignore some secondary effects of secondary effects of computation. They are taken into account in advance (by diagonalization and generalization) and thus they are related to prevention.

It is interesting to note that some scientists may intuitively 'feel' that Ackermann's function provides a model of human thinking of 'everything' for a particular situation. The above mentioned makeclear program shows that this intuition can be presented in terms of symbiosis of the information included in a particular situation. Note that the above macro-procedure (Figure 6) only simplifies the computation of thinking of 'everything'. In order to illustrate this simplification of the computation we may mention that, as it can be checked, the computational trace of the value for ack(3,2) using standard definition shows (see [8]) that the value ack(1,1) is computed twenty-two times for obtaining the result of ack(3,2). This is not the case for $f_3(2)$ simplified computation. It is however necessary to understand that the overall complexity of this situation remains the same since, in order to be able to 'simplify' (i.e., to define the above macro-procedure), we already need to have available Ackermann's function equivalent sequence of f_i. In other words, the principle and effectiveness of 'thinking of everything' are globally unaffected. The simplification concerns only focusing on one particular local level defined by the two values a and b instantiating Ackerman's variables. Of course, the macro-procedure is general, but for a and b given, it generates only the finite sequence of primitive functions $f_0, f_1, \dots, f_a$.

This makes explicit that 'thinking of everything' keeps its order of complexity after applying our simplification. Systems requiring a simultaneous handling of prevention and control factors such as information security systems or strategic planning in flexible environments are practical examples of a problem requesting to think of 'everything' (see [16], [10]).

F. Systemic recursion

In the previous sections, we have presented what can be seen as a recursive structure. In a recursive structure there is an obvious so-called base step element (for example, 0 in NAT). We shall speak about systemic recursion when a system is defined or constructed recursively, but there is no such an obvious 'base step element'. It is mostly the case when all the parts of the system may themselves be considered as symbiotic systems. In other words, in systemic recursion, symbiotic constructors are also systems, as will be illustrated by the following example – a simple one though it illustrates the complexity of systemic recursion:

Consider a method M that is recursively defined in terms of a finite number of complex symbiotically dependent procedures $R_1, R_2, \dots, R_n$. Then, a systemic equation for such a method can be represented as follows:

$$M = R_1 + R_2 + \dots + R_n + M.$$

Of course, rules $R_1, R_2, \dots, R_n$ may themselves be recursive procedures calling M as well. This emphasizes the difference between systemic recursion and linear-like, tree-like or network-like representations.

The above construction of Ackermann's function and our particular disentangling of its computation by a primitive recursive macro-procedure allow us to consider it as a model for a particular kind of Pulsation. The notion of oscillation, defined in the next section, provides an informal background for the notion of Pulsation as described in Section V.

IV. OSCILLATION

In scientific fields, an obvious basic paradigm, for a given problem, is looking for ideas that possibly lead to a solution. This behavior reflects the belief that the following formula is valid

$\forall$ Problem $\exists$ Idea Leads_to_a_solution(Idea,Problem).

We shall call this formulation: "first paradigm."

However, another and rather unusual (except in Physics) paradigm is to find an idea that provides a solution for all problems. We shall show how Ackermann's function provides a model for this second paradigm. Similarly to Ackermann's function, in a sense, it is a kind of 'thinking of everything'. First, however, let us express this paradigm by the formula

$\exists$ Idea $\forall$ Problem Leads_to_a_solution(Idea,Problem).

We shall call this formulation: "second paradigm."

The difference between these two formulas lies in the fact that, in this second case, the 'Idea' obtained is unique, while in the first formula each problem can use its own Idea.

We call **oscillation** this approach of *symbiotic* switching between the two above paradigms. It corresponds to a representation of a one-level creation process.

The oscillation may be performed in the following way. We start to consider a large variety of problems for which we try to find an idea for a general solution to all of them. This solution needs to be open to the need of a further improvement. We shall, in the next section, introduce Pulsation as being a particular kind of such an improvement,

V. PULSATION

The above sections will help us explaining how Ackermann's function enables us to formally specify the notion of **Pulsation**, i.e., a particular kind of 'evolutive improvement'. This is interesting not only from the point of view of building particular deductive theories for unknown domains but also for understanding the difference between revolution, innovation and evolutive improvement in this building process.

Let us consider a potentially infinitely incomplete theory. In unknown environments this may be seen as a framework for potentially infinitely incomplete theories. Building a deductive theory becomes then a process of *suitable completions* of a particular initial theory T_0. We shall say that this theory T_0 is **practically complete** when it formalizes solutions of practical problems that have been met so far. It implicitly means also, as mentioned in Section II.D, that a non ambiguous specific model is available. Since the theory is potentially incomplete, sooner or later we shall meet a problem that cannot be solved in the framework of T_0. In the vocabulary of scientific discovery we may say that we need a conceptual switch (a new axiom or a set of axioms) that *completes* T_0. Note that we speak here about *completion* and

- not about a *revolution* - which would mean in a sense rejecting T_0
- not about a *innovation* - which may simply amount to a particular reformulation of T_0, not necessarily coherent with T_0.

This completion T_1 has to contain T_0 and thus it must be coherent with T_0. However, since a new conceptual switch guarantees that T_1 is more powerful than T_0, we consider this particular kind of completion as a suitable model for one step of *improvement* in our search for suitable completions. Since we consider here a potentially infinitely incomplete theory, we can then see **Pulsation** as an infinite sequence of theories T_0, T_1, …, T_n, … . In this sequence, T_{i+1} completes and is coherent with T_i for all i = 0, 1, 2, …

We have seen that, in the infinite sequence from which Ackermann's function is built, the function f_1 relies on (is coherent with) f_0, and f_{i+1} relies on f_i for each i. We can therefore see that Ackermann's function really provides a model for evolutive improvement (or progress in Bacon's sense [2]) and we understand it as being different from revolution and innovation.

Let us now come back to our notion of Pulsation. We have seen that, in the informally specified notion of oscillation, we switch coherently between two paradigms. In our interpretation, the second paradigm, i.e.,

$$\exists \text{ Idea } \forall \text{ Problem Leads\_to\_a\_solution(Idea,Problem)}$$

represents the idea of Ackermann's function and the first paradigm, i.e.,

$$\forall \text{ Problem } \exists \text{ Idea Leads\_to\_a\_solution(Idea,Problem).}$$

represents particular primitive recursive functions from which Ackermann's function is constructed. In the definition of Ackermann's function we have seen that

$$f_{i+1}(0) = f_i(1).$$

Analogously, we shall state that the sequence of completing theories can be written as:

$$T_{i+1} = T_i + A_{i+1},$$

where A_{i+1} is an axiom or a set of axioms representing the conceptual switch that enables solving the problem unsolvable in T_i and solvable in T_{i+1}.

Let us stress the fact that by Pulsation we understand an infinite sequence of theories T_0, T_1, …, T_n, T_{n+1}, … with the just above mentioned property: It does not reduce to one particular step in this sequence. This means that pulsative systems are formalized progressively and potentially indefinitely.

We have seen above that Ackermann's function is also a model for symbiotic consideration of prevention and control.

Let us return therefore to the construction of Ackermann's function. We could see that, with respect to our requirement to obtain a non-primitive recursive function, f_0 must be defined in a way that guarantees the non-primitive recursion of the final composition of the constructed infinite sequence. Indeed, if f_0 were a constant, for instance 3 (which would mean that $f_0(n) = 3$ for all n), the resulting infinite composition would also be the constant 3. This means that, even though f_0 is the first function of this infinite construction, since it must be defined as a symbiotic part of the final composition, prevention and control factors must already be present in this function.

We can thus see that Ackermann's function provides in fact a model for the Pulsation that intends and guarantees symbiotic handling prevention and control already from the start. In other words, prevention and control are present already in T_0.

VI. Pulsation and Cartesian Intuitionism

In the previous part, we have introduced the notions of symbiosis, systemic recursion and Pulsation.

We have seen that symbiosis is different from compositions that are not separation-sensitive. Usually, systems that are not separation-sensitive are considered as modular also in the case of interdependency. In modular interdependent systems, the parts, when separated, preserve their essential properties. This is not the case for symbiotic parts of a symbiotic system.

We have also seen that recursive systems are different from systems that allow linear-like, tree-like or network-like representation.

It is somewhat obvious that the paradigm of Pulsation is different from what is understood as a process of evolution that tends to preserve only strongest 'individuals'. A similar kind of evolution can be recognized in self-organized systems. Edward de Bono [3], one of recognized experts of practically exploitable creativity and innovation, has characterized such a self-organizing system by the fact that "an idea may be logical and even obvious in hindsight but invisible to logic" of externally organized systems.

Conversely, in externally organized systems, any idea which is logical in hindsight must be accessible to logic in the first place. This last assertion is true for the classical systems but not for our SRPS. Pulsative systems are externally organized by human creators and developers. However, in pulsative systems no conceptual switch can be considered as logical in hindsight. This follows from the fact that the axiom (or system of axioms) A_{i+1} that extends a theory T_i is logically independent from the previously constructed theories. In consequence, A_{i+1} cannot be logically explained in T_i.

Moreover, the main problems of Pulsation are construction of systems and development of completion-like procedures for these systems. The decision procedures are secondary and dependent on the developed construction and completion-like procedures. In Pulsation, all 'individuals' collaborate symbiotically towards one goal that is informally specified from the start.

This means that SRPS are complementary to non-recursive systems that are usually considered in science or business. In order to capture the essential difference between the paradigms implicitly present in standard science and the complementary SRPS, we call these paradigms Newtonian and Cartesian Intuitionism, respectively.

The main difference between Newtonian and Cartesian paradigms is easily perceptible from comments pronounced by Newton and Descartes themselves.

In a letter to Robert Hooke, Newton wrote: "If I have seen further (than you and Descartes) it is by standing upon the shoulders of Giants."

Newtonian science can be seen as established on logic of sequential 'observational' research. In a little more systemic way, we can thus describe the Newtonian way by a sequence of advancements built one upon the other from a 'beginning' until an 'end'. We say that this research is observational since, at each step of advancement, it does not require that the previous results are fully recreated. They are only observed externally and adopted as true. This is a model of what means "standing upon the shoulders of Giants."

Descartes wrote his first rule in the *Discourse on the Method of Rightly Conducting the Reason, and Seeking Truth in the Sciences* [4] in a following way: "The first was never to accept anything for true which I did not obviously know to be such; that is to say, carefully to avoid precipitancy and prejudice, and to comprise nothing more in my judgement than what was presented to my mind so clearly and distinctly as to exclude all ground of doubt."

This formulation could be looked upon as being similar to Newton's except that Newton expresses the utmost confidence in the 'giants' while Descartes wants to check, or rather re-create everything by himself before accepting a new knowledge. Indeed, Descartes always recreated all the knowledge useful to him when it had been previously obtained by someone else. This means that, if ever standing on 'giants' shoulders' took place, it had to be very carefully checked in order justify any extension of it.

Descartes justifies these possible extensions by stating that they have to lead to some obvious truth obtained by what he called 'intuition'. He describes what this 'intuition' is in his *Rules for the direction of the mind* (*Regulae ad directionem ingenii* [5]). When examining his definition from a recursive systemic point of view, we get hints that intuition, for Descartes, is a symbiotic, possibly recursive, composition. The process by which these hints are shown to be reasonable is complex and explained in detail in [8].

The same thing is expressed by Descartes in a little more complicated way by saying that "beginnings … can be persuaded well only by the knowledge of all the things that follow later; and that these things which follow cannot be understood well, if we do not remember all those that precede them." [4], p. 797. Note that our description of a Pulsation, in Section V. above, looks like an explicatory paraphrasing of Descartes' way of speech. From a more formal systemic point of view, we may state that the demarcation of a notion is not the initial stage, as it is the case in the Newtonian paradigm, but the final stage of its formation.

We thus see that Descartes' work, as we present it, contains a basis for SRPS systemic research. We introduce therefore Cartesian Intuitionism as a paradigm complementary to the Newtonian one. For us, Cartesian Intuitionism is nothing but a systemic science relevant to SRPS. In [9] we provide a more detailed comparison between Cartesian Intuitionism and Newtonian approach in the framework of Program Synthesis (PS). PS is a basic problem to be solved in the technological vision described in the next section.

VII. A Pulsative Technological Vision

In the previous parts, we have introduced the basic notions for a rough understanding of SRPS. Further work is necessary to provide a deep understanding of systemic emergence, i.e., the 'how' behind Pulsation. Symbiosis and recursion of parts of a system are reason why SRPS cannot be well understood externally. It is necessary to study them in the framework of a concrete creational referential system. In Computer Science, automation of recursive Programs Synthesis in Incomplete Domains via Inductive Theorem Proving (PSIDITP) already provides a usable experimental creational referential system. A formalization of this problem as well as a description of one particular approach built on systemic science of SRPS can be found in [9]. This approach is called Constructive Matching Methodology (*CMM*).

Let us make precise here what we call a methodology in a technological framework: Given a non-trivial goal, its solution relies on a fully formalized 'algorithmic' description of all problems that arise in achieving this goal. In this context, this special description is what is called a methodology (for the solutions of these problems). In other words, a methodology is a full 'know-how' for successfully achieving the given goal.

From the point of view of Pulsation presented here, it is interesting to note that the goal of *CMM* is to build a program synthesis system ('Idea') providing a 'Solution' to the problem of program construction in incomplete theories.

We thus globally work with the second paradigm. However, in our everyday research (which means to acquire fruitful experiences enabling to build relevant knowledge), we work locally with the first paradigm while keeping in mind the second paradigm. This means that we mentally oscillate between two paradigms. The second paradigm presents a global vision and the direction of the solution we seek and, to make this goal achievable, we perform our everyday work in the framework of the first paradigm following nevertheless the direction imposed by the second paradigm. It is important to note that we are still at the level 0 of pulsative development of *CMM*. In other words, we work on defining a powerful primitive recursive f_0 with respect to the overall goal of resulting non-primitive recursive SRPS for *CMM*. This means that level 0 has already required several decades of research and many useful results not known in PSIDITP were obtained so far. A full bibliography of these results can be found in [12]. We have to underline here that, obviously, an informal version of the Pulsation model was used from the start of our research. Recently only we formalized it enough to be presented in [1]. Our experimental implementation in [7] reflects this pulsative feature of our research.

In the long term, an expected success of the mentioned approach to PSIDITP provides fundamentals for a pulsative technological vision that may roughly be described by three contributions of PSIDITP. Indeed, PSIDITP seems to be a way how robots, in the future, will be able to

- formalize recursively unknown domains (e.g., in space research) handling perfectly control, rigor and evolutive improvement;
- perform experiments necessary for finding such suitable formalizations;
- program themselves autonomously with the help of the formalizations found.

Formalizing an unknown domain is a progressive exploration aimed at acquiring experiences – through experiments – that lead to facts enabling some progress in the formalization of this domain.

Of course, a successful achievement of this technological vision will require other tools than the ones presented in [9], [11]. New tools developed in Machine Learning, Big Data, Computational Creativity will certainly be also necessary. It is even quite possible that some of the necessary tools will appear in the future, born from pluridisciplinary cooperation and from yet unknown scientific fields developed in the course of research in PSIDITP.

VIII. Pulsation, SRPS and Ancient's systems

The above presented construction and role of Ackermann's function as a model for infinite pulsation provides a very good sieve through which it is interesting to study or revise the systemic foundations of Ancient civilizations. Eternity, Timelessness and Progress are three essential themes upon which grow these ancient foundations (see [6], [2]). In modern interpretations these notions are still embedded within philosophical opinions. In our opinion, the pulsation model provides another point of view, a more mathematically oriented one.

We use a feature central to Ackermann's function, namely its representations by a specific infinite sequence of different and non-trivial functions, which constitutes a computable representation of *eternity*. Each of these functions plays an important role throughout the progressive growth of the sequence.

Timelessness might be represented by the fact that each f_i contains in a sense all the previous f_j (for $j < i$) and thus there is no obsolescence.

Finally, the fact that f_{i+1} represents a sort of a conceptual switch extending the potential power and action of f_i and of all previous f_j (for $j < i$) is a rigorous representation of *progress*, though perhaps a bit limited one.

This means that Ackermann's function seems to be a very good start for a more mathematically rigorous model of the three ancient symbiotic notions of Eternity, Timelessness and Progress.

This kind of thinking leads us to detect possible roots of our SRPS approach in ancient philosophy: It could quite be a particular "déjà-vu" of what has been understood in Ancient times as Universal Mathematics. In this case, it follows that a systematic study of the systemic links between all these ancient systems of thought and SRPS might bring many new ideas and technological visions for modern Science in general and for the development of secure dynamic evolutive systems in particular.

IX. Conclusion and Future Work

Recognizing the symbiotic character of systems is vitally important, namely for security reasons and for preserving the essential properties of systems that are designed for showing no decline. We have shown that recursion is able to handle symbiotic information in systems and participates in their rigorous control and prevention. For real-world applications it seems therefore useful not only to recognize but also to develop symbiotic systems whenever conceptual switches are needed. With respect to the incompleteness of human knowledge, such a development must be strategically planned in order to avoid hindering future systems evolutions. We have therefore defined a non-trivial model, called Pulsation, for creation of symbiotic recursive systems. This model shows the essential features expected for a smooth evolution of human knowledge and for the design of ambitious real-world applications, namely

- possibility of infinite evolution (i.e., no decline)
- possible coherence of new results with previously developed systems, i.e., no rejection,
- rigorous security and prevention handling (i.e., no accidents).

In this way, the paper substantially completes and refines our definitions of Cartesian Intuitionism and Symbiotic Recursive Pulsative Systems introduced in our previous work.

We have mentioned in the paper the pragmatic and structural character of the notions developed. We have also mentioned a real-world application in which these notions are embodied. Their dynamics, i.e., their algorithmic descriptions started to be described in [13], where we provide a simple illustration a particular feature of Pulsation. Next, we plan to extend and generalize our experience acquired mostly in the process of the application mentioned in this paper.

Because of their rich potential, we foresee that Cartesian Intuitionism and Symbiotic Recursive Pulsative Systems will play an important role in the innovation process without needing to compete with standard Newtonian paradigms. However, once the complementary and non-competitive character of symbiotic systems is well understood, together with their rich potential, Cartesian Intuitionism and Symbiotic Recursive Pulsative Systems will certainly be highly exploited regardless of the particular ways of thinking they require.

Acknowledgment

We thank to Michèle Sebag and Yannis Manoussakis for a moral support.

References

[1] M. Franova, Y. Kodratoff, "A Model of Pulsation for Evolutive Formalizing Incomplete Intelligent Systems," in L. van Moergestel, G. Goncalves, S. Kim, C. Leon, (eds), INTELLI 2017, The Sixth International Conference on Intelligent Systems and Applications, ISBN: 978-1-61208-576-0, 2017, pp. 1 - 6.

[2] F. Bacon, The Advancement of Learning, Kessinger Publishing Company, 1994.

[3] E. de Bono, Serious Creativity: Using the Power of Lateral Thinking to Create New Ideas, HarperCollins, 1992.

[4] R. Descartes, Philosophican works, Oeuvres philosophiques (3 vol.), Edition de F. Alquié. T. 1, Classiques Garnier, Bordas, 1988.

[5] R. Descartes, *Rules for the direction of the mind,* "Regulae ad directionem ingenii," in R. Descartes: Oeuvres complètes (11 vol.), Edition Adam et Tannery. T. 10, Vrin, Paris, pp. 359-469, 1996.

[6] I. Franco, Rites and beliefs of Eternity (Rites et croyances d'éternité), Pygmalion - Gérard Watelet, 1993.

[7] M. Franova, Precomas 0.3 User Guide, Rapport de Recherche No.524, L.R.I., Université de Paris-Sud, Orsay, France, October, 1989.

[8] M. Franova, Formal Creativity, Method and Use – Conception of Complex "Informatics" Systems and Epistemological Patent (Créativité Formelle : Méthode et Pratique - Conception des systèmes « informatiques » complexes et Brevet Épistémologique), Publibook, 2008.

[9] M. Franova, "Cartesian versus Newtonian Paradigms for Recursive Program Synthesis," International Journal on Advances in Systems and Measurements, vol. 7, no 3&4, pp. 209-222, 2014.

[10] M. Franova, D. Hutter, and Y. Kodratoff, "Algorithmic Conceptualization of Tools for Proving by Induction 'Unwinding' Theorems – A Case Study," Rap. de Rech. N° 1587, L.R.I., Université de Paris-Sud, France, Mai 2016.

[11] M. Franova and Y. Kodratoff, "Cartesian Handling Informal Specifications in Incomplete Frameworks," Proc. INTELLI 2016, The Fifth International Conference on Intelligent Systems and Applications, pp. 100-107, 2016.

[12] M. Franova: List of publications (retrieved 2018.05.25) https://sites.google.com/site/martafranovacnrs/publications .

[13] M. Franova, Y. Kodratoff, "Cartesian Systemic Emergence - Tackling Underspecified Notions in Incomplete Domains," in. O. Chernavskaya, K. Miwa (eds.), Proc. of COGNITIVE 2018 : The Tenth International Conference on Advanced Cognitive Technologies and Applications, ISBN: 978-1-61208-609-5, pp. 1-6, 2018.

[14] D. Gries, The Science of Programming, Springer-Verlag, New York, 1981.

[15] S. C. Kleene, Introduction to Meta-Mathematics, North-Holland, 1980.

[16] H. Mantel, A Uniform Framework for the Formal Specification and Verification of Information Flow Security, PhD thesis, Universitty of Saarlandes, 2003.

[17] R. W. Ritchie, "Classes of predictably computable functions," Trans. Amer. Math. Soc. 106, pp. 139-173, 1963.

[18] E. W. Weisstein, Young Girl-Old Woman Illusion, From MathWorld - A Wolfram Web Resource. http://mathworld.wolfram.com/YoungGirl-OldWomanIllusion.html (retrieved 2018.05.25).

[19] A. Yasuhara, Recursive Function Theory and Logic, Academic Press, New York, 1971.

[20] Euclid, The Thirteen Books of the Elements, vol. I, II, III, Dover Publications Inc., Édition, 1956.

From Compartments to Agents via Fuzzy Models - Modeling and Analysis of Complex Behavior of Physiological Systems

Sebastian Meszyński[1], Oleksandr Sokolov[1], Aleksandra Mreła[2]
[1]Faculty of Physics, Astronomy, and Informatics
Nicolaus Copernicus University
Toruń, Poland
[2]Kujawy and Pomorze University in Bydgoszcz, Faculty of Technology

email: sebcio@fizyka.umk.pl, osokolov@fizyka.umk.pl, a.mrela@kpsw.edu.pl

***Abstract* – It is well-known that analytical modeling and computer simulation of the physiological systems is a complex problem with a great number of variables, equations, and non-linear relations. There are several approaches for such modeling. One of them is so-called compartment modeling. Each compartment is assumed to be a homogeneous entity, within which modeled entities are equivalent. Another approach is multi-agent modeling, which consists of creating agents with a more complex structure in comparing with the compartment and a more complex logic of behavior and communication. In the paper, we decided that before building a multi-agent model to try to describe the communication of the system elements using the rules of the class 'state-action' and transferring such relations to the properties of the agents. In order to analyze the dynamics of the behavior of the multi-agent system, the matrix description method was proposed. As an example, we investigate in the article the different models of the insulin-glucose physiological system.**

Keywords – multiagent system; agent; linear algebra; matrix.

I. INTRODUCTION

The models used for modeling of any kinds of physical phenomena are the tools utilized to obtain an answer to questions concerning the tested system, without the need for performing the actual experiment. Among the variety of models, i.e., psychological, word or physical models, there are also mathematical models whose relations observed in the system are described by mathematical formulas. The possibility to perform such experiments is called simulation (lat. simulare – simulate). It is a cheap and safe alternative or a complement to experiments with the system.

The quality of simulation's results depends entirely on the quality of the model. Fundamentally, there are two approaches to building a model representing a particular system. The first type of approach is based on the knowledge taken from literature or experience of experts in each domain and could be used for building more and more precise description of the investigated phenomenon (more complex models are generated). The second one is based on observation of the phenomenon and its behavior on one level of description (using similar agents) and after that building the model and identification of parameters (agent-based approach).

The created model in both approaches needs to be described in a handy form, especially if one wants to analyze it with the use of digital machines. Having the model built, it is necessary to verify the correctness of obtained results. The credibility of the results provided by the model can be acquired using verification or validation.

This paper focuses on the use of MAS a multi-agent system (MAS) for the modeling of the insulin-glucose system responsible for the blood glucose homeostasis. Even by designing the simplest model based on the multi-agent paradigm, one must rely on a complex analysis of interactions between agents. For this reason, there is not one general formalism of description of these interactions, which would additionally allow an easy analysis of the functioning of such a multi-agent system. In most cases, the used approaches are chosen depending on the category of the problem that is solved by the system. If MAS was designed to address the issues of game theory, then this formalism would be used to analyze the multi-agent system. When MAS was created for optimization problems, these problems will be used to analyze this system [9,14]. What is presented in this paper is a demonstration of the use of two modeling techniques for the general description of a multi-agent system. On one hand, the theory of compartment models has been used to describe the interaction between the different body regions, called compartments. On the other hand, graph theory introduces a general and universal tool for describing the interaction between beings that can represent any mathematical or physical concept. Combining these two techniques allows us to describe the interaction between agents in MAS in two ways. Firstly, it could help to describe the dynamics of the entire multi-agent system, showing the connections between agents, their behavior, and the ability to investigate the whole system. Secondly, it makes possible to include in the same formalism the information associated with each agent. This should be understood as the ability to get information, about which behavior is implemented in the body of the agent, which is used to communicate with the environment, and which is only the internal behavior of the agent. One can also get information, about which agents are receivers of the messages and, which are senders of those messages.

The proposed approach allows describing MAS in two complexity scales – the system as a whole and the agent and its impact on the system. The proposed advantage is a technique, which can extend the behavior of compartments using fuzzy logic and make their behavior more complex by describing their internal structure using agent description [1]. We illustrate the MAS description and communication of the glucose homeostasis. The selected analytical model (Stolwijk-Hardy model [11]) was converted to MAS in a lossless fashion. As a result, individual members of this model became the determinants of behavior of individual agents, and in addition, the analysis of such model was maintained by compartmental methods.

The structure of the paper is following. Section II describes the compartment models and their applications in physiologic systems. In Section III we introduce the fuzzy logic models and illustrate them in the example of the insulin-glucose system. This model will be a base of agent's behavior in the next section.

Section IV gives a short introduction to multi-agent systems and draws attention to components of agents and their communication standards. In the next section, the matrix representation of MAS is proposed. Section VI illustrates the authors' approach with two simple examples. The conclusion and references summarize the article.

II. Compartment systems

The concept of a compartment is not unambiguous and may represent different features depending on what is being discussed. Usually, the compartment describes the structures of a living organism characterized by similar properties in relation to the test substance [5]. The compartments may be either separate areas of the body or substances. In the first case, the compartment can be considered an organ or intercellular space, in the latter case, it may represent blood

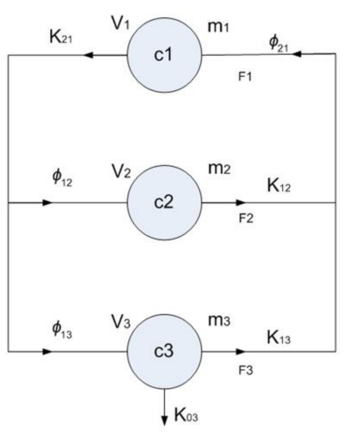


Figure 1. An example compartment model (the model has three compartments).

plasma. If the test substance is in the biological system in several areas of the organism and it is possible to determine its movement between these areas and changing its concentration, then such areas can be considered as separate compartments (Figure1). The description of the transport of substances in the body is considered for the mass balance (or concentration) in individual compartments. The main reason for using this description is the ability to determine the time course of substance exchange streams between compartments and between the body and the environment.

The considered models are constructed assuming the constant volume of compartments. In this case, the mass of substance in the compartment is proportional to its concentration. The equations describing the transport can be summarized for both mass and concentration. The transition from the first description to the second one consists only in dividing both sides of the equation by a constant factor, which is the volume of the compartment.

Considering the *i*-th compartment, the mass of substances in this compartment can be changed as a result of the algebraic summation of the input and output streams of this compartment:

$$\frac{dm_i(t)}{dt} = \sum_{\substack{j=0 \\ j\neq i}} \phi_{ij}(t) - \sum_{\substack{j=0 \\ j\neq i}} \phi_{ji}(t)\,,\; i = 1,2,\dots,n \quad (1)$$

where $\phi_{ij}(t)$ denotes an input stream flowing from the *j*-th compartment to the *i*-th compartment; $\phi_{ji}(t)$- the output stream flowing from the i-th compartment to the j -th compartment; *n* - a number of model compartments.

By grouping the streams in the right order: resultant interconnection exchange rates, the elimination stream, the dosing flow and considering the biological availability F_i equation (1) for the mass balance of the *i*-th compartment can have the form:

$$\frac{dm_i(t)}{dt} = -\sum_{\substack{j=1 \\ j\neq i}} \phi_{ij}(t) + \sum_{\substack{j=1 \\ j\neq i}} \phi_{ji}(t) - \phi_{0i}(t) + F_i d_i(t), \quad i,j = 1,2,\dots,n \quad (2)$$

The subscript 0 denotes the connection of a given compartment with the external environment, where $\phi_{i0}(t) = F_i d_i(t)$ is the substance dosing flow to the *i*-th compartment, and ϕ_{oi} is a stream of elimination flowing out of the *i*-th compartment. Biological availability F_i is a fraction of the given dose $d_i(t)$, e.g., a medicinal substance that has been absorbed into the *i*-th compartment. This parameter meets the condition $0 \leq F_i \leq 1$. In cases of linear pharmacokinetics, this equation acquires the features of a differential linear equation. The following types of linear pharmacokinetic models can be distinguished:

• full-time or part-time,
• without delay or with a delay.

In linear stationary models, the streams are of the donor type, i.e., they form a function of mass $m_j(t)$ in the compartment, from which they flow and are proportional to this mass:

$$\phi_{ij}(t) = k_{ij} m_j(t)$$

where k_{ij} is the constant speed of exchange.
The counterpart of equation (2) is a linear differential equation with constant coefficients:

$$\frac{dm_i(t)}{dt} = -\sum_{\substack{j=1 \\ j\neq i}} k_{ij} m_i(t) + \sum_{\substack{j=1 \\ j\neq i}} k_{ji} m_j(t) - k_{0i} m_i(t) + F_i d_i(t) i, j = 1,2,\dots,n \quad (3)$$

where: k_{ji}, k_{ij}, k_{0i} are constant inter-compartment exchange rates and elimination velocities, respectively.

In the case of the two-compartment model (Figure 2),

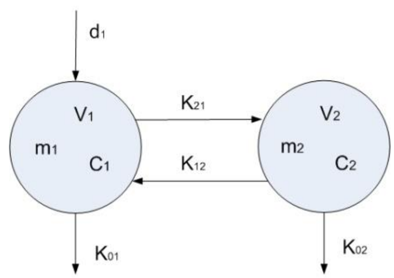


Figure 2. An example of the two-compartment model.

the equations describing the mass balance take the form:

$$\frac{dm_1}{dt} = -(k_{01} + k_{21}) m_1(t) + k_{12} m_2(t) + F_1 d_1(t),$$

$$\frac{dm_2}{dt} = k_{21} m_1(t) - (k_{02} + k_{12}) m_2(t). \quad (4)$$

In the general case of a multi-compartment system, the system of equations looks like this:

$$\frac{d\mathbf{m}(t)}{dt} = \mathbf{A}\mathbf{m}(t) + \mathbf{F}\mathbf{d}(t) \quad (5)$$
$$\mathbf{c} = \mathbf{C}\mathbf{m}(t) \quad (6)$$

satisfying the initial conditions:

$$\mathbf{m}(t = 0) = \mathbf{m}(0)$$

The solution of equation (3) is defined by the formula:

$$\mathbf{m}(t) = e^{\mathbf{A}t}\mathbf{m}(0) + \int_0^t e^{\mathbf{A}(t-\tau)} \mathbf{F}\mathbf{d}(\tau) d\tau \quad (7)$$

Below there is an example of such a model for the regulation of the insulin-glucose system [13] (Bergman and Cobelli model)

$$\frac{dg}{dt} = -[a_1 + x]g + a_1 G_B,$$
$$\frac{dx}{dt} = -a_2 x + a_3 [i - I_B], \quad (8)$$
$$\frac{di}{dt} = a_4 [g - a_5]^+ t - a_6 [i - I_B],$$

in which the test substances are insulin and glucose, which levels we denote by i and g, respectively. Depending on the parameters of the model (8), simulations, as well as analyzes of the normal state and diseases of diabetes can be made.

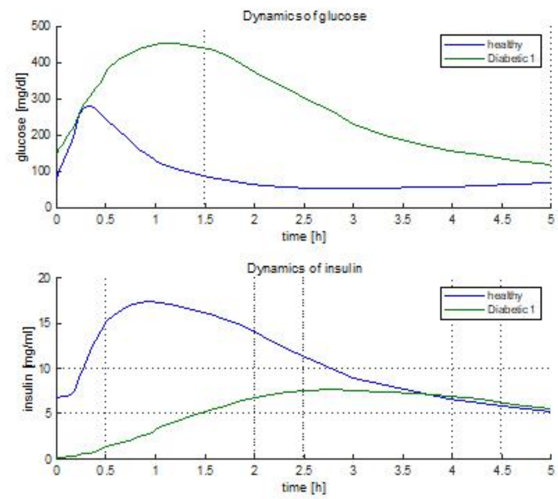


Figure 3. An example of insulin-glucose dynamics for a healthy person and person with diabetic I

Analysis of the dynamics of a multi-compartment system is given by the state equations (5), the output equations (6) consists in examining the features of solving (7), controllability and stability by analyzing the eigenvalues of the fundamental matrix A. Usually, such analysis is carried out by using the Laplace transform.

The optimization of the systems (5), (6) depends on calculating the given dose $\mathbf{d}(t)$, by means of which the quality functional for the solution (7) is minimized, for example, the preservation of a given mass change program $\mathbf{m}(t) \rightarrow \mathbf{m}^*(t)$. Well-known methods of control theory allow a solution, in which regulators are used, which are "tuned" through the selection of appropriate settings. Thus, we apply the classic methods of analysis and optimization.

The advantages of the compartmental model can be described in a uniform description of all organs that communicate with each other by means of substance transfer. The disadvantages of such a description can be the inability to describe hierarchical models and the difficulty in estimating the values of all parameters in the case of a non-linear model.

III. Fuzzy Logic Models

Biological systems can be described using a quantitative or qualitative way. Unfortunately, the quantitative approach using the compartmental description and linear models (5) and (6) does not bring the desired results. Due to the non-linearity of physiological processes, the complexity and uncertainty of the biological system and the occurrence of delays and measurement deviations, it is difficult to build a model with correctly selected parameters. These specific features of the biological system force the use of a different type of modeling.

Qualitative models are relatively simple and require basic knowledge about the modeled system to correctly map it. Quantitative models simulate the analyzed system very precisely, but they require knowledge of accurate kinetic data, which are sometimes not known satisfactorily or are simply missing. There is one more approach to building models - it is a semi-quantitative approach that uses tools such as fuzzy systems and fuzzy logic [4]. They allow the description of the system in a satisfactory way even when the data related to the kinetics of the biological process are not complete. This incompleteness can also be seen as the external and internal variability of the biological system. The assumptions that should be met by the modern model of the biological system are presented in the article by Parker [8], in which the author lists the elements necessary for such a model like that: prognostic skills within the input-output process, the ability to perform calculations using the Internet for control and optimization.

The use of this type of modeling comes to combining functional blocks in a proper way (Figure 4) and implementing the methods used in a given simulation technique in their structure.

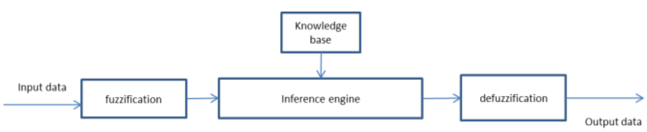


Figure 4. Fuzzy logic model.

The general principle of creating a system consists in breaking the model into several functional blocks:

- input block – responsible for entering heterogeneous input data into the system and their conversion to the internal format (for fuzzy logic this process is a fuzzyfication),
- processing block – processing of the received information using a knowledge base for this purpose - rules of fuzzy logic (for fuzzy logic these will be inference procedures),
- output block – the transformation of the calculation result generated in the processing module to the format of the user-understandable model or to the format used by the rest of the model (for fuzzy logic this process is defuzzyfication).

To create a system, it is worth to use graph models, in which the nodes are individual elements, and the edges are a substance and information transferred from one element to another.

For each of the individual organs of the insulin-glucose system, we create sets of rules describing its behavior according to the knowledge of its internal state and parameter states in other organs.

The rule has a structure of recursive relationships that combine the state at a discrete time k and k+1:

$$\textit{If } X_{1k} = NB \textit{ and } X_{2k} = ZE \textit{ and} \ldots X_{jk} = NS \textit{ Then } X_{i,k+1}=PS, \tag{9}$$

where X_{jk} are variables that characterize the content of substances in the body j, and NB-negative big, NS-negative small, ZE-zero, PS-positive small, ... are linguistic variables that use the membership functions to determine the full range of the value of the respective substance.

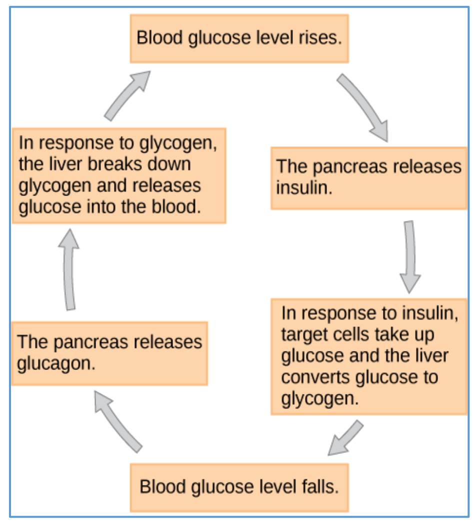


Figure 5. Dynamic of glucose transportation.

Type (9) rules allow creating dynamic relationships, for example, for the insulin-glucose system (Figure 5). Each block of the system contains a model of fuzzy logic.

One of the main advantages of the fuzzy model is a very simple way of describing the communication of organs of the physiological system in the form of rules (9) and the selection of parameters as membership functions of linguistic variables. The disadvantage of such a system is the difficulty of analyzing and creating the optimal control of such a system.

IV. Multi-Agent Systems

We present here the basic ideas concerning multi-agent systems.

A. *Concept of the multi-agent system and agent*

MASs are complex systems of agents communicating and cooperating with each other. This construction of the systems enables solving problems of a diffuse or complex calculation. In the studies applying multi-agent systems, the concept of an agent is presented as an autonomous object having the initiative of action based on the observation of the environment, in which it is located. It also has the ability to use the resources of the environment and the motivation to solve the problem it has to face. Such definition of the agent forces it to have inputs called sensors (through which it will be able to receive signals from the environment) and effectors, which can be used to influence the surrounding environment. The most important task of the agent is to decide, which of the possible courses of action is best, at the time of acquired knowledge about the problem, in order to achieve the goal.

The issue „agent" is wide and diverse. Nowadays, the term is so broadly used that is best described as comprising a heterogeneous body of research and development [2,7]. Different communities refer to it in various ways. Some scientists will characterize agents as initiatives and reactivity of objects; others emphasize independent learning and communication skills. What can also be invoked is the characteristic that unifies modeling agent the most – it is their decentralization. An extensive discussion of multi-agent systems can be found in positions [15,12].

In contrast to the dynamic system or actions based on models, MAS does not have a special place of centralization where the dynamics of the system is fixed. In addition, the global behavior of the whole system is determined based on the individual behavior of all agents. Each agent has its own internal behavior as a set of rules and behaviors for interacting with the environment and other agents. This description creates a dynamic interaction of agents based on the rules.

In many situations, there is a doubt linked to the lack of understanding of the philosophy of using multi-agent systems and returning toward object-oriented programming. What is characteristic of multi-agent systems can be presented in the following subparagraphs:

- Agents possess internal awareness and defined goals to be achieved. The goals can, but do not have to be identical to the objectives of the other agents who are in the same environment. In such case, information obtained from another agent can be considered only if it is coincident with its own objective.
- An agent is a dynamic instance, which adapts its activity to instantaneous changes in the environment and has certain fixed parameters and characteristics only for it that do not change regardless of the extent of the changes observed in the environment.
- Each agent possesses at least one strand, which is responsible for its behaviorism.

The general difference between instance of an agent and an object lies in the fact that the object has variables that change, while the agent variables can be changed only when the agent accepts the request of the sender to change the value of a variable in an immediate way or after the act of negotiation.

B. *Communication in multi-agent system*

In an environment where there is more than one agent, there must be a mechanism for the exchange of information between the environment and the agent, and between agents. Communication mechanisms are essential for the agents grouped in structures that facilitate co-operation so that they could achieve their goals. Since the multi-agent environments [6,10] are dynamic environments, it is necessary to introduce a mechanism that would allow for informing the agents of the existence of other participants in the system. The literature [3] distinguishes the following approaches:

- Yellow pages, where agent can place information about services it provides,
- White pages – the list of all agents in the environment,
- Broker – intercessory agent.

To create a message and then send it to another agent in such a way they can receive it and understand it, it is necessary to define a common communication language. It should be noted that the communication language, which is independent of the field, is separated from the language of messages content. Among the communication standards the most popular ones include:

- KQML (Knowledge Query and Manipulation Language),
- ACL (Agent Communication Language).

Among the examples of the language of message content, the following should be distinguished:

- KIF (Knowledge Interchange Format),
- FIPA standards:
  - SL (Semantic Language),
  - CCL (Content-Language).

Having a tool for communication, agents can communicate with each other to achieve a common or an opposing goal. In the first case, we have to deal with the concept of co-operation, in the second case - with the competition concept. As a rule, multi-agent systems are designed to solve complex problems, in which agents have control (or can observe) only over a certain part of the environment (Figure 1). If MAS can solve the problem, the agent has to have knowledge and control over the entire environment. To do this, the agents are organized into a structure, in which they can interact with each other. Interactions between structures and agents are supposed to bring them benefits. Each agent has its preferences for the state, in which environment it should be (this is its goal). To describe this preference, the concept of utility ν was introduced, which causes the state of alignment of the environment Ω due to the agent's preferences.

$$\nu : \Omega \rightarrow \Re \qquad (10)$$

The environment that corresponds to preferences of the agent will have greater utility value (in other words: the agent will "feel better").

V. Matrix description of the multi-agent system

The paper key objective is to propose a modeling glucose-insulin paradigm in the form of MAS starting with a mathematical description and finishing the implementation of the program. This solution shows how we can implement features of agents for both the macro and micro processes in homeostasis of glycemia. Moreover, at the same time, we can allow operating on two scales: organs and cells scale. This approach results in a new quality of information. To describe the multi-agent system, the authors used the approach describing compartment modeling and using the rationality of graph theory. This approach simplifies the interpretation of what is happening in the multi-agent system, therefore, the behaviors of individual agents and their influence on other agents can be easily identified in the considered system.

This approach simplifies the interpretation of what is happening in the multi-agent system, therefore, the behaviors of individual agents and their influence on other agents in the considered system can be easily identified.

The analysis of MAS is a difficult task to implement due to the existence of the asynchronous relationships between agents occurring in the system. Additionally, each agent, which takes an active part in MAS has at least two behaviors: the first one is to receive incoming messages from other agents, and the other one is used by it to send the information to the chosen agent. By verification of the model, one can understand two aspects. The first one concerns information about the acceptable range of internal parameters of the model, which guarantees the stability of the model for the incoming information/extortion from outside. The second aspect concerns the range of input set, which ensures the correct stability and expected representation of the behavior of the modeled system.

We propose to describe MAS by using a comparison of network connections between the agents with the connections between vertices forming a graph. Nomenclature of the vertices is extended by the occurrence of behaviors that identifies the agent's behavior. In this perspective of the problem, the graph, which describes the interactions between agents with their associated behaviors is obtained. The assumptions are:

- Behaviors implemented in a given agent create a set of behaviors for the agent, which is a subset of behaviors occurring in the multi-agent system:

$$\sum A \in \Phi \tag{11}$$

$$\Phi \subseteq \Omega \tag{12}$$

  where:
  A - represents some behavior of an agent, Φ - represents a set of behaviors of a given agent, Ω - represents a set of behaviors of a multi-agent system.
- Agents who present the same behaviors are not identical with each other. It causes independent actions in the terms of time and each agent using the same behavior performs them in various time slots.
- Graph A=(V,E); |V|=n, |E|=m represents MAS basing on the assumption that:
  - n: number of graph vertices (number of agents),
  - m : number of behaviors appearing in MAS.
- Adjacency matrix $K \in M(n \times n; N)$ is defined in such a way that value in *i*-th line and in *j*-th column equals:
  - 0: if there is no communication between agents (no connection),
  - 1: if there is communication between agents (connection).

  Whereby:
  - k_{ii} represents the cyclical route of agent *i*-th,
  - k_{ij} represents the route from agent *i*-th to agent *j*-th.
- The sum of the same behaviors is one behavior:

$$\sum_i A_{1i} = A_1 \tag{13}$$

- Behavioral matrix $A \in M(n \times n; B)$ (where B designates the set of behaviors within the scope of the multi-agent system) is defined in such a way that a value in i-th line corresponds to behavior responsible for communication between agent *i*-th and agent *j*-th, whereby:
  - Behavior A_{ii} represents internal behavior (cyclical) of agent *i*-th,
  - Behavior A_{ij} represents information exchange from agent *i*-th to agent *j*-th.

Taking the above assumptions into consideration, it is possible to describe MAS with the use of matrix equation:

$$A^T K + D = \Phi \tag{14}$$

where:
A^T is the transpose of a matrix of agents' behaviors; K is a matrix of connections between agents; D is a matrix of agents' internal behaviors; Φ is a matrix representing the multi-agent system.

Analysis of the above equation will be presented on examples of multi-agent system. Both examples will rely on a different number of behaviors occurring in the multi-agent system.

VI. Examples

In this paragraph, authors demonstrate examples of the use of a matrix to describe MAS and to select unknown behavior.

A. The example of two-agent description based on the matrix representation

Let us consider the multi-agent system, where two agents A1 and A2 have predefined behaviors, and A11 and A22 are internal behaviors and A12 and A21 are external behaviors (Figure 6).

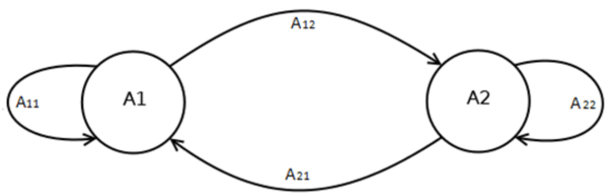


Figure 6. Two-agent system.

For the following example, adequate matrixes will be defined:

$$A = \begin{bmatrix} A_{21} - A_{12} & A_{12} \\ A_{21} & A_{12} - A_{21} \end{bmatrix} \quad (15)$$

$$A^T = \begin{bmatrix} A_{21} - A_{12} & A_{21} \\ A_{12} & A_{12} - A_{21} \end{bmatrix} \quad (16)$$

$$K = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (17)$$

$$D = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \quad (18)$$

Substituting to equation (15) we obtain a representation of MAS in the form of:

$$[\Phi] = \begin{bmatrix} A_{11} + A_{21} & A_{21} - A_{12} \\ A_{12} - A_{21} & A_{22} + A_{12} \end{bmatrix} \quad (19)$$

Conducting a detailed analysis of the matrix Φ we receive information about:

- First minor (φ_1) of a matrix Φ represents internal and incoming behaviors to agent $A_{1:}$

$$\varphi_1 = A_{11} + A_{21} \quad (20)$$

- Second minor (φ_2) of a matrix Φ represents behaviors of data exchange between agents A_1 and $A_{2:}$

$$\varphi_2 = A_{21} - A_{12} \quad (21)$$

- Third minor (φ_3) of a matrix Φ represents data exchange between agents A_1 and $A_{2:}$

$$\varphi_3 = A_{12} - A_{21} \quad (13)$$

- Fourth minor (φ_4) of a matrix Φ represents internal and incoming behaviors to agent $A_{2:}$

$$\varphi_4 = A_{22} + A_{12} \quad (22)$$

- Trace of a matrix represents behaviors occurring in the multi-agent system:

$$Tr[\Phi] = A_{11} + A_{21} + A_{22} + A_{12} \quad (23)$$

The examples were designed to show the application of (14) to describe MAS and the equivalence with the use of a graph. Description using matrixes is helpful in such a way that, in a compact form, it contains a representation of the dynamics of the multi-agent system. It is not relevant what type of behaviors are written using matrix A. That is why the authors consider this record as universal. The results matrix Φ contains much information, from which one can restore the functioning of the multi-agent system, basing solely on the content of individual cells of the matrix. Individual cells φ_i make it possible to obtain information on what types of behavior are present in the agent - whether they are its own internal behaviors (e.g., A_{11}) or behaviors associated with taking or receiving information to/from another agent (e.g., A_{21}). Additionally, the sum of the behavior of a given line (e.g., $\varphi_1 + \varphi_2$) is interpreted as the behavior occurring in the agent (e.g., for A_1). The results matrix can also determine whether, in the multi-agent system, there is at least one bidirectional communication between agents. To verify whether in MAS the exchange of information occurs, it is necessary to check whether the following identity is met:

$$Tr[\Phi] = \sum_i \varphi_i \quad (24)$$

To verify the above relationship, the examples discussed earlier can be used:

$$A_{11} + A_{21} + A_{22} + A_{12} = A_{11} + A_{21} + A_{22} + A_{12} \Leftrightarrow Tr[\Phi] = \sum_i \varphi_i \quad (25)$$

B. The example of matrix representation for identification of desired behavior

The experiment is quite specific. This uniqueness is based on the use of the matrix record, introduced in Section III, to determine unknown behavior in the multi-agent system. The experiment was based on a two-agent representation of the glucose homeostasis system. The first agent represents the entire mechanism of normoglycemia in the case of type 1 diabetic patient. The second agent represents insulin delivery in the form of external administration (Figure 7). The purpose of this experiment is to define the behavior responsible for sending "information" from Agent A1 to Agent A2 so that the dose of insulin delivered contributes to the metabolism of glucose.

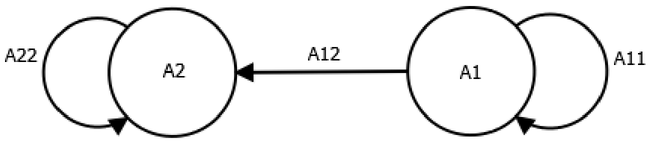


Figure 7. Diagram of MAS for the experiment.

Based on the concepts introduced in the section above, we can define the appropriate arrays, and so the matrix A:

$$A = \begin{bmatrix} -A_{12} & A_{12} \\ 0 & A_{12} \end{bmatrix} \quad (26)$$

matrix K:

$$K = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad (27)$$

matrix D:

$$D = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \quad (28)$$

The matrix of MAS is defined by the corresponding relation between the previously mentioned matrices so that the system matrix is:

$$\Phi = \begin{bmatrix} A_{11} & -A_{12} \\ 0 & A_{22} + A_{12} \end{bmatrix} \quad (29)$$

The trace of the matrix:

$$Tr\phi = A_{22} + A_{11} + A_{12} \quad (30)$$

In this particular case, the meaning of the individual behavior is as follows:

- Behavior A_{11} is responsible for the insulin production that will eventually be introduced into the system. This behavior may also represent a buffer that stores a certain amount of insulin.
- Behavior A_{22} represents all the phenomena occurring in the glycemic homeostasis system, along with the ways of insulin utilization.
- Behavior A_{12} is responsible for the exchange of information (from agent A1 to agent A2) - this behavior should be determined.

The purpose here is to define the behavior A12 in such a way as to ensure insulin levels of φ_{A2}=7 [uIU/ml] for Agent A2. Below, a procedure to achieve our goal is presented:

1. Simulation for the conditions specified for a person with type 1 diabetes (without insulin infusion) (Figure 10).

2. Transform the pattern (22) into a form that allows us to calculate the desired behavior. In this case, we get:

$$A_{12} = \varphi_{A2} - A_{22} \quad (31)$$

3. Perform curve fitting procedure (Figure 14) to the points obtained. This procedure was performed in MATLAB environment using the “fctool” command. The fit was done using a linear function. The following form of function is given:

$$f(A_{12}) = -0{,}0914t + 6{,}14 \quad (32)$$

4. The last step was to implement the equation described in (24) into the body of the insulin dispensing agent. The simulation was started and a comparative analysis of data from the insulin-free model and from the model, in which the found behavior A_{12}.

Below are the following drawings corresponding to the mentioned above points.

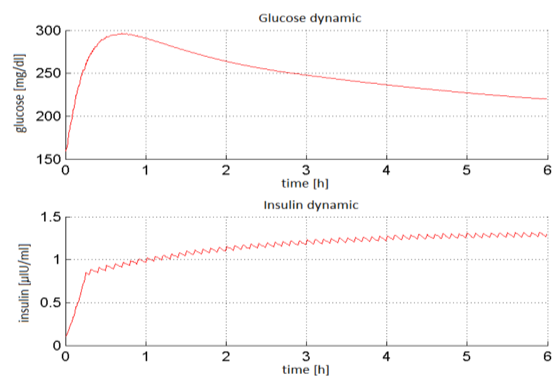


Figure 8. Simulation result for a person with type 1 diabetes - without insulin.

As can be deduced from Figure 10, the concept of using a matrix description to identify unknown behaviors is the most appropriate approach.

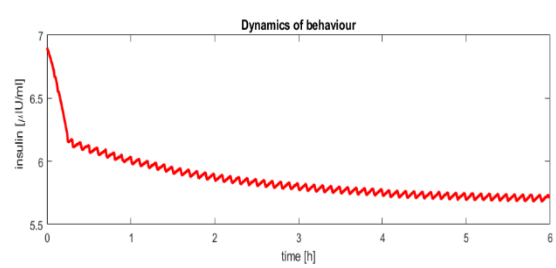


Figure 9. Chart for variability of behavior A_{12}.

Using (22), it is possible to select unknown behavior in such a way that the preset value can be maintained throughout the system under consideration. By focusing on the selected part of matrix φ, there is an opportunity to declare such an unknown behavior that will result each value from the agent the minor describes. This is the second case presented in this experiment. As a result of matching A_{12}, it has become possible to maintain insulin levels of 7 [μIU/ml] by the agent A2.

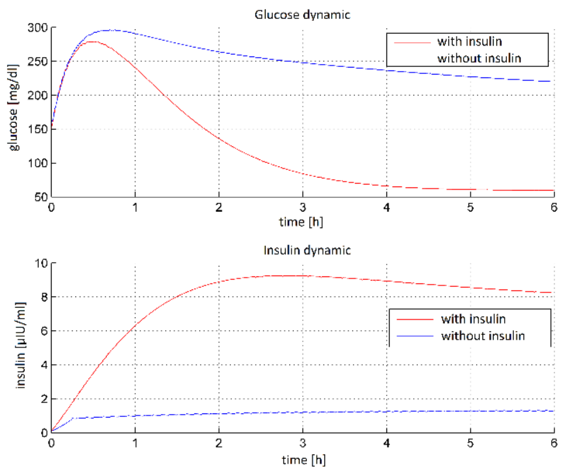


Figure 10. Simulation results for two cases: without insulin (blue curve), including the behavior of insulin dosing into the multiple agent system (red curve).

Of course, the quality of the curve fitting to the measurement points (Figure 8 and 9) directly affects the quality of the results generated by the multi-agent system.

VII. CONCLUSIONS

In this paper, we consider the problem of investigation of the complex biological system using compartment, fuzzy and multi-agent approaches.

We performed the analysis of MAS with the use of graph theory and matrix calculus. This approach can help us analyze the operation of such system in two ways: quantitative and qualitative ones. The use of matrix record enables performance of analysis of the internal multi-agent system involving the assignment of behaviors to particular agents. External analysis of MAS with the use of introduced record allows the description of the relation between agents and selection of such unknown behavior of an agent, which will meet the intended purpose or criterion implemented by the multi-agent system. In the second example, it is shown how using matrix equation allows finding the desired behavior of multi-agent system. For the general case, in which the agents (and the multi-agent system) process several volumes, each of these factors must be represented by a separate graph of accurate dependency. Generally speaking, each value can represent a different graph of connections between agents, and agents can have different numbers and behaviors intended to process these values. Matrix equation (15) proposed by authors, will be the subject of further work towards stability study of the multi-agent system.

REFERENCES

[1] S. Meszyński, and O. Sokolov, “Modeling of Complex Multiagent Behavior Using Matrix Representation,” INTELLI 2017, The Sixth International Conference on Intelligent Systems and Applications, pp.32-37, July 23-27, 2017 - Nice, France, https://www.iaria.org/conferences2017/ProgramINTELLI17.html.

[2] P. Agre, “Computational theories of interaction and agency,” Mit Press, 1996.

[3] F. Bellifemine, A. Poggi, and G. Rimassa, “Developing multi-agent systems with JADE,” International Workshop on Agent Theories, Architectures, and Languages, Springer Berlin Heidelberg, 2000.

[4] W. J. Bosl, “Systems biology by the rules: hybrid intelligent systems for pathway modeling and discovery,” BMC systems biology, 1(1):13, 2007.

[5] C. Cobelli, G. Toffolo, and E. Ferrannini, “A model of glucose kinetics and their control by insulin, compartmental and noncompartmental approaches,” Mathematical biosciences, 72(2):291–315, 1984.

[6] A. Helleboogh, G. Vizzari, A. Uhrmacher, and F. Michel, “Modeling dynamic environments in multi-agent simulation,” Autonomous Agents and Multi-Agent Systems 14.1: 87-116, 2007.

[7] H. S. Nwana, “Software agents: an overview,” The Knowledge Engineering Review, 1996, 11(3), pp. 205–244. doi: 10.1017/S026988890000789X.

[8] R. S. Parker and F. J. Doyle, “Control-relevant modeling in drug delivery,” Advanced drug delivery reviews, 48(2):211–228, 2001.

[9] S. Parson and M. Wooldridge, “Game Theory and Decision Theory in Multi-Agent Systems,” Kluwer Academic Publisher, 2000.

[10] T. Seth and U. Wilensky, “NetLogo: Design and implementation of a multi-agent modeling environment,” Proceedings of agent, vol. 2004.

[11] S. Soylu, K. Danişman, I.E. Saçu, and M. Alçi, “Closed-loop control of blood glucose level in type-1 diabetics: A simulation study,” Electrical and Electronics Engineering (ELECO), 2013 8th International Conference on. IEEE, 2013.

[12] P. Stone and M. Veloso, “Multiagent systems: A survey from a machine learning perspective,” Autonomous Robots 8.3: pp. 345-383, 2000.

[13] G. Toffolo, R.N. Bergman, D.T. Finegood, C. R. Bowden, and C. Cobelli, “Quantitative estimation of beta cell sensitivity to glucose in the intact organism: a minimal model of insulin kinetics in the dog,” Diabetes, 29(12):979–990, 1980.

[14] P. J. Wangerman and F. R. Stengel, “Optimization and Coordination of Multiagent Systems Using Principled Negotiation,” Journal of guidance, control, and dynamics, vol. 22, No. 1, 1999.

[15] M. Wooldridge, “An introduction to multiagent systems,” John Wiley& Sons, 2009.

Study of the Appropriation of Groupware in the Context of Remote Collaborative Design

Aspectuality via time, occurrence and changes

Samia Ben Rajeb
Université Libre de Bruxelles
Brussels, Belgium
email: samia.ben.rajeb@ulb.ac.be

Shima Shirkhodaei
University of Liege
Liege, Belgium
shima.shirkhodaei@ulg.ac.be

Pierre Leclercq
University of Liege
Liege, Belgium
pierre.leclercq@ulg.ac.be

***Abstract*— This article presents a methodology of analyzing the appropriation of groupware in the context of collaborative design via the idea of aspectuality (punctual, iterative, durable, inchoative, and terminative), well known in the field of Greimassien semiotics. This idea guides us in the definition and the categorization of the modes of appropriation of tools, as well as the passage from one tool to another during a collaborative activity. To do this, our research focuses on the study of an innovative device, associating two tools for remote synchronous collaborative design: HIS (Space Hybrid Ideation of Hybridlab), and SkeSha (Sketch system of sharing from LUCID-ULg). These two tools enable the annotation of graphical objects in real time. HIS enables immersion into the interior of a virtual representation of a designed space, the other (SketSha) enables the possibility to share and to act on the 2D documents. In our experiments, these two tools were associated to form a system enabling two groups of student designers to work together remotely and in real time. Two questions came up in this original experimental situation: the first concerns the singularity of each tool and the second touches on the degree of compatibility of the two devices making up a system for synchronous exchange and collaboration. To answer these questions, we will describe the experimental protocol put in place in the simultaneous use of these two tools. Then, we will present our methodology of analyzing the data based on self-analysis as well as the qualitative and quantitative treatment of the data put to work in the experiments. Finally, we explain in detail in which manner the two devices are complimentary and can be articulated in the preliminary phases of architectural design.**

Keywords-component; groupware; appropriation and tools uses; semiotcs; aspectuality; computer human interaction.

I. INTRODUCTION

The rapid evolution of operating technologies in the field of collaborative design raises not only the question of the singular use of each tool, but also the influence of their association in this activity and during the action. In this context, we present here an analysis of the modes of appropriation of an innovative device, associating two tools to instrument distant and synchronic collaborative design. In this paper, we explore a previous study, presented in [1]. The two tools are the Hybrid Ideation Space (HIS), developed at the Hybridlab, a laboratory of University of Montreal [2, 3], and the Sketch Sharing system (SketSha), developed at LUCID, a laboratory of University of Liege [4, 5]. Both are based on the notation of graphic artifacts in real time. One (HIS) allows immersion in the interior of a virtual representation of a conceived space, the other (SketSha) makes it possible to share and act on 2D documents. In the experiment, these two tools were associated to allow two groups of student designers from University de Liege and School of Architecture of Nancy to collaborate, under the direction of the HybridLab team. Two questions emerge from this original experimental situation: the first concerns the singular implementation of each tool and the second concerns the degree of programmatic compatibility in the use of a device, which integrates various tools for exchange and synchronic collaboration. To answer these questions, Section II first describes the experimental protocol implemented in the simultaneous usage of these two tools. In Sections III and IV, we present the methodology of data analysis based on the notion of aspectuality (punctual, iterative, durative, inchoative and terminative), well known in the field of Greimasian Semiotics. This notion guides us to the definition of determining categories to explain the switching from one tool to the other during the collaborative activity.

Our approach focuses on the methodological aspect to enable the analysis of complex collective activities involving new technologies. This is why our state of art only concerns the methods and shows why we have resorted to aspectuality to address this kind of problem (see Section III).

Based on quantitative and qualitative analyses, Section V will show that the degree of familiarization of users with the new technologies is a determining factor to characterize the issues and the limits of this superposition of tools. Finally, we will also detail to what extent these two complementary devices can be articulated in order to support preliminary phases of architectural design.

II. FRAMEWORK AND RESEARCH QUESTIONS

This research is part of collaboration between the LUCID laboratory at the University of Liège and Hybridlab at the University of Montreal. Both HIS and SketSha devices, developed in in the universities of Liege (Belgium) and Montreal (Canada), were enabled to instrument collaborative design.

SketSha software enables real-time sharing of drawings and annotations, via a digital tablet horizontally placed in front of the designer, drawn by using an electronic pen during a remote meeting. Images, PDF, DXF drawings or other documents can be imported and made available to all partners of the project. These documents are shared on the basis of a stack of semi-transparent tracing paper that users can annotate, store, superimpose or manipulate in real time.

HIS is a device based on an immersive system for placing various remote users within their graphic representation, their sketched freehand drawings and three-dimensional models "on which they interact by manual and digital actions". This complex device mainly consists of two parts: (1) a digital tablet placed horizontally showing a 2D image of the project. The image is chosen by the designer and depicts the localization of the project intervention. This image allows drawing and annotation via an electronic pen; (2) a piece of canvas that is hung vertically to close the work space in which the designers act. The same image that is pre-treated to provide users with a 360 ° view of the inside the project can be projected on its surface. This projection helps designers immerse themselves in real time in their sketches while drawings appear on the tablet in front of them.

An experiment involving these two devices to design a project was set up (Figure 1). Two groups of designer (students of University of Liege and the School of Architecture of Nancy), who were geographically distant, worked for about 3 hours. The synchronous use of HIS and SketSha at this collaborative meeting involved two virtual work spaces that share a resembling feature, namely the sharing of graphic documents in real time on the digital table between the users taking part in the meeting from two geographically distant offices. However, these two devices are distinguished by the HIS-device's immersive dimension.

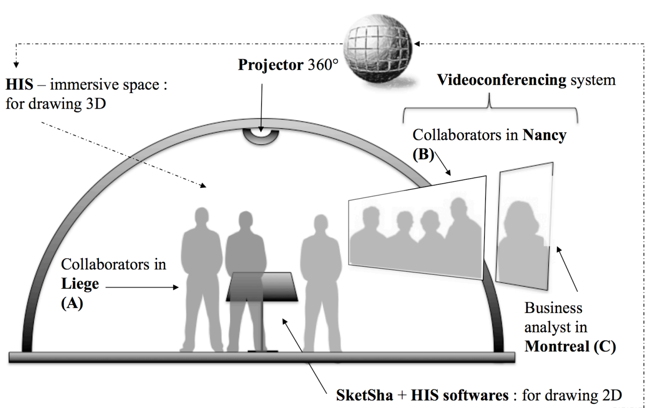


Figure 1. Context of experiment.

Therefore, our first research question relates to the activity of actors in each work space called (Work HIS and Work SketSha). Thus, we will study the "duration" and "occurrence" of the two main activities of actors were studied, namely designing and being able to look in both work spaces. Our second research question concerns the modes of switching from one work space to another. Our hypothesis is based on the existence of two types of switching used by the actors: (1) switching between Work HIS and Work SketSha, (2) switching between 2D and 3D.

It should be noted that although the HIS requires physical precedence of some immersive space throughout the meeting, the mode of the presence of the immersive space for the meeting depended primarily on the activities of users and how they made this immersive space (from 2D to 3D) real. On the other hand, it was necessary to compare these remarks with collective operations involved in this collaborative architectural design. This parallelism enabled us to notice the specific particularity of time used for each tool during a collaborative session. Once we determined the decisive moments of the two types of switching, we noticed the specificity of these changeovers and then analyzed them from the point of view of the aspectualization defined in the field of linguistics and semiotics.

III. Methodological Position

The question that we pose is: how can the ideas related to the notion of aspectuality help us describe the complex collective activities and enable us to specify the methods of changing from one immersive work space to another work space? In fact, other scientific fields have taken an interest in the analysis of collective activities. For example, in sociology, the question has been asked in terms of the organization of actors' roles in a team [6]; or in terms of recognition, personal satisfaction and confidence among the different members of a team [7]. In cognitive ergonomics, the questions are centered on the interactions between partners, on the synchronization of the collective activity of design and on the cognitive aspects [8]. When the activities involve new technologies, one finds oneself in the scientific fields of CSCW (computer supported cooperative work). Moreover there are different points of view to analyze this kind of complex activity [9, 10, 11]:

1) the point of view of the physical aspects of the work: this point of view is only interested in the ergonomic and physical aspect of the space in which the designer works. We speak of the physical space with its acoustic and thermal properties, gestuality, movements, postures, etc.

2) the point of view of the affect is concerned with the psychological or emotional aspects of the designers. This aspect measures the subjective feelings of the designers in relation to their surroundings and their collaborator. Thus, it deals with hierarchical relations and feelings of confidence that unite the different members of a team;

3) The cognitive point of view looks at the cognitive aspects of the design process that are linked to the situation, the actors and the subject in question. In this case, the conscience of the group, the intermediary objects and the shared reference are parameters to be considered to study these situations;

4) The organizational point of view's objective is to define the modalities of assistance to the situations of group work or to help in managing group-design documents.

Our paper proposes another point of view, which tackles the collaborative design activity involving new technologies: semiotics. The reference to the notion of aspectuality in linguistics and in Greimasian semiotics [12, 13] helps us to address the question of the appropriation of these two tools considering time, occurrence and switching. The definition of Holt [14], p. 6, is one of the first attempts to define aspect. According to Holt, aspect concerns "different ways of conceiving the flow of process". The nucleus of this definition remains unchanged. The notion of aspect is currently used in linguistics as a grammatical category that expresses the subject representation of a process denoted by a verb [15] p. 53. Thus, a verb, an adjective or a noun can be analyzed in terms of aspectualization. For example negotiation or decision-making are aspectualized substantives, insofar as the first is considered as an unfinished act and the second as an act already completed. For Bertrand [16], "aspect modulates the semantic content of the predicate, whether it is in past, present or future". Via this notion of aspectuality it is possible, for example, to address the issue of the progress of a process otherwise than by time. For example, if the aspect is taken in terms of time, it is called "punctual" or "durative". The aspect can be described as "terminative" when it is approached from the point of view of its completion and "inchoate" when it is intended to be the beginning. Here, the process is not only related to time but also concerning the state of its switching (see Section V). This specification in the synchronous use of two tools, supporting collaborative design in an architectural design project, led to the issue of proportion via the aspectuality relative to time, occurrence and switching.

Our methodology is therefore based on this concept of aspectuality with the aim of analyzing quantitatively and qualitatively complementary data from this experiment. A coding scheme was defined for the transcription of a user's activities before the semiotic analysis of the processed data. In concrete terms, it is a matter of leaning of the three fundamental to elements of aspectuality (time, occurrence, and switching) to analyze the method of appropriation of the system and to evaluate more precisely the stakes, the limits and the perspectives of each single modality ("drawing" and "looking") and complex ("collective operations of design) during the use of these two tools. Thus an adjustment practice was put forward including speech, drawings and looks. The manners were specified in the two tools have been appropriated by the different participants / designers. But before going directly to the presentation of the results, we propose to clarify the context and the protocol of this experiment.

IV. EXPERIMETAL PROTOCOL

A. Experimentation

Our protocol is part of a defined framework, which involves the following parameters.

- Role of the actors participating in the experiment: the participants are formed with a sponsor (Actor C is in Montreal and is also the moderator of the session) and 7 designers divided into 2 groups (group A is in Liège and group B is in Metz);
- Hierarchical relations between the designers and their expertise related to the use of the tool: these designers are trained by experts (represented by teachers who are used to using the two tools) and novices (represented by students who have already used Sketsha several times but have worked with HIS only one time);
- Training of the designers: Group A includes 3 actors of whom 2 are students in Master’s Engineer-Architect and their teacher at the University of Liège (A1, A2, A3) while Group B includes 4 actors of whom 2 are architecture students, their teacher in architecture and another teacher specialized in ergonomic psychology (B1, B2,B3, B4):
- Problem of design, pointed out by the sponsor, in relation to the rearrangement of a library: to solve this problem graphic elements (images of the interior space treated in 360° able to be projected on the HIS screen and some plans as well as simple pictures taken in the space can be used and annotated directly on SketSha) were made available to the designers and shared.

All the actors were first invited to use the whole system (HIS and Sketsha simultaneously) at least one time. This experience 0 gave them the opportunity to take in hand the tool and to exercise in a completely different context before doing the experiment concerned by this study.

All the geographically remote actors work in the same kind of environment that associate HIS and SketSha. The designers share the graphic annotations and exchange orally via the video conference in real time. The problem the 2 teams must work on consists to more precisely rethink the library of the future starting from an existing site and a real context. The designers formed from groups A and B are led to think about the possible uses of current spaces of the library in order to propose a rearrangement of the space better adapted to contemporary uses and new TIC technologies. Their work is about graphic documents that already exist on SketSha and pictures taken in the library and prepared to be visualized in HIS. On SketSha, 3 documents are shared: (1) a plan of the present floor in consideration with the furniture; (2) a plan of the present floor in consideration without the furniture; (3) a view of the ground and the insertion of the building on the site. On HIS, different views in human scale are projected, annotated and manipulated.

At the end of this experiment, 3 hours and 15 minutes of video recording had been taken by the Montreal team. To make the work of the researchers easier, this recording is made up of a juxtaposition of 4 views showing (Figure 2): a shot of all the traces made by SketSha (upper left), a view of the HIS environment in Montreal, appearing in the same way as on all the other sites (upper right), a view from the webcam in Metz (bottom left) and a view from the webcam in Liège (bottom right).

Only 2 hours were treated in the framework of our study. The first quarter hour of set up was deleted and an hour and the end of the work seance during which the designers could no longer communicate because it was interrupted by recurring moments of bugs caused by the video-conference system.

At the end of the experiment, we had a semi-open interview with the designers in order to get their spontaneous feedback. In the context of our study, only the designers from group A were questioned because we were geographically in the same place. We did not try to separate the designers during the interview because our research did not focus on the designers' individual use. On the other hand, we were more interested in the appropriation of the system by a group of users participating in a collaborative activity.

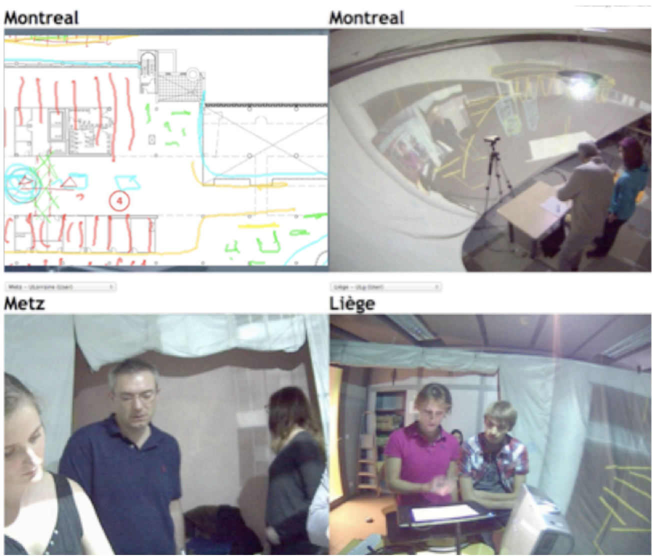


Figure 2. Juxtaposition of 4 views: graphic elements produced by the designers and views from Montreal, Metz and Liège sites.

B. Methods of data processing: from self-analysis to video processing

In the logic of complementary data and inspired by the field of cognitive ergonomics, our methodology of data processing was divided into two steps. First, we applied the method of ergonomics with group A with the aim of getting their feelings (tiredness, concentration, stress, annoyance, discouragement, enthusiasm, etc.) and their feedback in relation to their use of the tool, but also in relation to their appropriation of the whole system. To do this, we used the semi-directed interviews made at the end of the experiment to construct our chart of self-analysis. Secondly, we processed the video by making cuts in relation to a coding scheme, which was specified by our state of the art, but also by that which was cleared and highlighted in the framework of the self-analysis with group A.

1) Protocol of the self-analysis

Lasting about 3 hours, several steps made up the self-analysis done with group A in the context of our study:

- we reminded the designers the context of the experiment and the main steps, which made up their design exercise;
- then we explained the modalities of the self-analysis that consists of commenting on 10 short sequences of about 4 minutes from the video recorded during the experiment. These 10 sequences were selected in a way to cover the group of phases, which made up their design exercise;
- in addition to the video placed on the table, we gave them a frame of reference made up based on the interview carried out with them following the experiment. This chart took into consideration the following points:
  - relationship between the actors,
  - evolution of the project over the time,
  - modalities of communication between the actors,
  - specific use of each tool,
  - appropriation of the system using the two tools when they were used simultaneously.

  An example of this frame was placed on a table in front of them, near the video (Figure 3). Opposite them, on the wall, were posted 10 examples of this chart representing each of the 10 sequences selected for this seance of self-analysis;

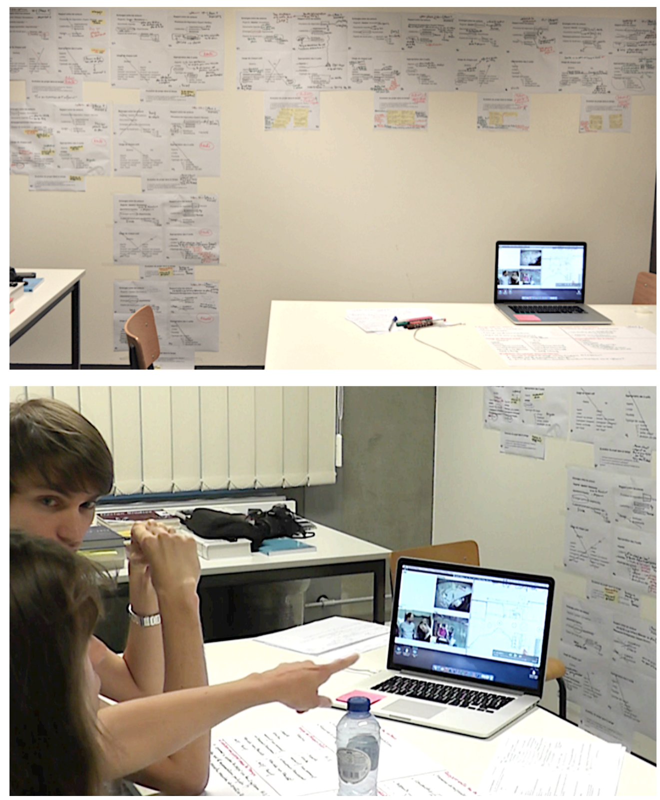


Figure 3. Self analysis set up.

- after looking at each sequence, the chart was collectively filled in with the researcher using pens (with different color codes related to the feelings of each actor) but also post-its so they had the possibility to add something if they think that an example, an explanation or an argument should be added to one or another element than the chart offered. The designers also have the possibility to go back while looking at the video in order to verify something that was said or one of their feelings;
- when all of the charts shown on the wall were filled in, the designers were asked to re-do a synthesis

relative to their individual or collective or collaborative activity on the tool, or their collective activity or interaction with the other group. In other words, it means firstly, to spot their concrete actions on this or that tool and, secondly, the moments where they think they could have collaborated with group B. Then we asked them to also spot moments when each group isolated themselves on the side without the possibility of negotiation or collaboration between the two groups;

- at the end of this exercise, a feedback on the use of HIS/SketSha was requested in relation to their experiment 0 by which they had learned for the first time to use the system.

At the end of this self-analysis, the participants highlighted the contribution of this kind of method which, according to them, enabled them to " show the other aspects of what had happened… it was while watching the video that I realized the impact of the tools on our communication over distance and our group work...".

2) Protocol of data processing via SketSha Replay

The coding done here concerns the video of the experiment. It enabled us afterwards to quantitatively analyze the data in addition to that which had been caught qualitatively during the self-analysis. For this coding, the data were treated via SketSha replay. This software, designed and developed in the LUCID laboratory, enables the coding of a video recorded the coding of a collaborative seance according to exclusive criteria. In our case, these criteria were specified thanks to elements that were highlighted by the designers during the self-analysis. The coding criteria centered most precisely on the three following criteria: "drawing", "looking", and "doing together", the objective being to spot participants' actions in the two work spaces, immersive and non-immersive, in relation to the use of HIS and SketSha. Out of the three main activities two types of categories emerged: simple and complex. The first takes into consideration the individual intervention of the users in the shared graphic space; the second was deduced from the collective activity of each of the two groups of collaborators (A and B). To do this, we first proceeded with a temporal cutting of the sequence (from 1 to 6) in relation to the different steps of the designing process. From this temporal cut, we then selected a sequence, which especially shows the switching from one tool to the other, as well as the use of the two.

3) Division into sequences.

This division remains nevertheless subjective even if it was validated during the self-analysis. It depends mostly on the objectives researched in the framework of this study. A sequence indicates, in our opinion, " a series of sequential choices forming a narrative unit that answers to a general problem by the actors during the designing process". Each sequence involves thus a beginning and an end marking the passage from one subject to another and/or the transformation from one state to another (and/or a proposal) but does not systematically end with a solution (and/or an answer to this proposition) [17], p.185.

In this way the work seance being studied here is cut into six sequences (Figure 4):

- Sequence 1 (45 minutes) – Understanding the request: this sequence is the longest in the design process. Now, the sponsor (C = commanditaire) explains his request and all the other actors try to undeerstand the objectives aimed at by the new project. The actors begin by visualizing on SketSha then on HIS all of the elements that they have been given. Then, they try to construct a common understanding of the plan that has been given to them by SketSha. After this, all of thee actors try to clarify the request in order to specify the new elements to be integrated in the project.
- Sequence 2 (6 minutes) – Increasing the space dedicated to reading: after discussion, the 2 groups of designers decide to increase the space dedicated to reading estimating that it is the first priority for the arrangement the future library.
- Sequence 3 (14 minutes) - Integrating the light: in negotiating the arrangement dedicated to reading, the designers decide to create two kinds of space (zones of conviviality and zones for reading) according to their proximity to openings.
- Sequence 4 (18 minutes) – Integrating new technologies: by trying to optimise the space, the designers question certain existing functions and thus decide to integrate more adequate new technologies to the future library.
- Sequence 5 (21 minutes) – Calling into question: the designers call into question the whole current program of the library and try to answer the question "What function to give to the future library"?
- Sequence 6 (12 minutes) – The first attempt to rearrance: after the intervention of the sponsor, the designers decide to work immediately on one of the main spaces of the present library having good light quality and whose facade has an non-standard shape: that of a bite.

In the context of this article, we have chosen a coding aimed at a particular segment in order to gather our quantitative data.

4) Choosing the segment.

To choose this segment, we first proceeded with a cut relating to the work spaces used by the actors during the design process. We based this on the verbalisation and the intention expressed by the actors when they asked to modify, explicitly, the work space to validate a point of view. We proposed this code for the entire length of the work meeting with the objective to tafe into account all the switches from one tool to another during the experiment, which is perfectly coherent with our objectives from the start. We also stressed the importance of the moments during which the designers did not work together: "Bug" moments caused by problems with the video conference and moments of "Logistic" management. Thus, this first cut was done according to the following criteria:

- "HIS" work space: here, the actors used the HIS device (by drawing in 2D on the digital tablet placed in front of them and looking at their interventions projected on the screen with 3D printing) for synchronous sharing of the documents, the discussions and evaluation of their proposals.
- "SketSha" work space: here, the actors used the SketSha software (by drawing in 2D on the digital tablet put in front of them) during the meeting.
- "Logistics" moment: all the moments when the actors communicate in order to adjust the problems concerning the logistics are coded as belonging to the logistic.
- "Bug" moment: it concerns technical and computing problems that caused the interruption of exchanges in the actors' communications.

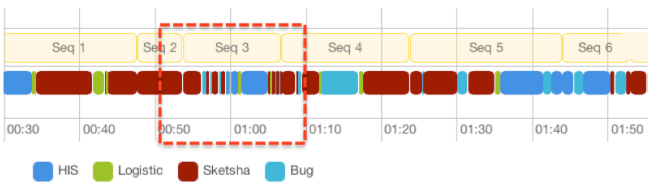


Figure 4. Progress in the design process and selection of the treated segment.

This first cut then enabled us to focus our analysis on the segment centered on sequence 3 that was characteristic of marking several switches between the two tools. In order to assure the coding precision of this sequence and thus decrease errors of interpretation, we included in this segment a bit of the sequence that preceded it and a bit that followed. Thus, on the temporal axis of the observed meeting, the segment we dealt with according to our coding scheme began at 50 minutes and ended at 1h10. Nevertheless, only the data that concerned sequence 3 were analyzed quantitatively, in the context of this article, in order to not alter our specific results to changes made during the sequence.

Thus, we proceeded with a second cut of this selected sequence in sub-sequences relating to the work space used, the objective being to more precisely observe the appropriation of each tool separately (SketSha / HIS), as well as the change from one to another (1 SketSha / 2 HIS / 3 SketSha). Like the first cut, we emphasized the indications relevant to the verbalization marking this change from one work-space to another (Figure 5):

- A2: "Is it important that one can switch on the HIS?"
- B1: "Shall we switch to SketSha?".

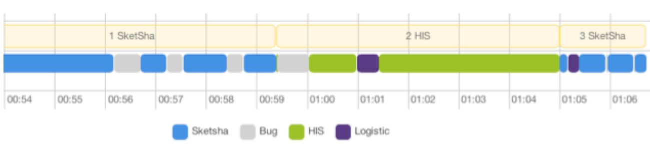


Figure 5. Progress of sequence 3 according to the designers' work spaces.

5) Coding the sequence 3 with SketSha Replay

Two types of codes were done for sequence 3 via SketSha Replay: the first (simple category) takes into count the individual intervention of each user in the shared graphic space ("Drawer" Category and "Watcher" category); the second, "Doing Together" comes from the collective activity of each of the two groups of collaborators (A and B).

a) Simple Category: drawing and watching

Drawing. This category involves three criteria (Figure 6):

- drawing SketSha: actors draw on SketSha;
- drawing HIS: either actors draw on the tablet (2D) or they draw on the immersive space (canvas gives a 3D effect);
- not drawing: the players do not draw.

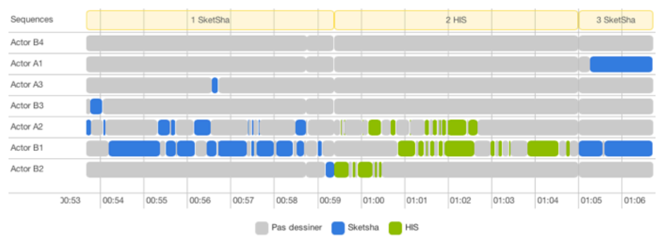


Figure 6. Drawing SketSha, drawing HIS and no drawing actions.

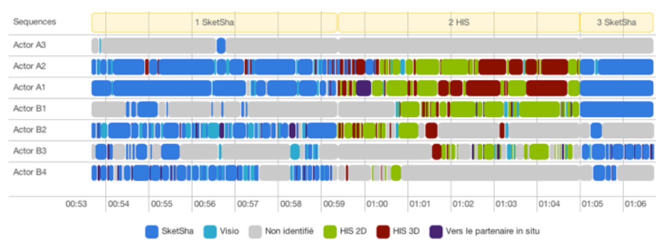


Figure 7. Looking SketSha, Visio, Unidentified looking, lookinh HIS 2D, HIS 3D and looking at another in situ actions.

Looking. This category involves six criteria (Figure 7):

- looking SketSha: actors look at and follow the documents on SketSha;
- looking HIS 2D: actors look at documents on HIS 2D;
- looking HIS 3D: actors look at the documents on the HIS in the immersive space;
- looking Visio: the actors make contact with their partners in inter-teams by looking at the videoconference;
- looking at the other in situ: actors see their partners in the same team;
- Unidentified looking: looks are not identified (e.g., out of sight for observer).

b) Category complex: collective operations of design

Processing of this category is to detect the different operations carried out by each of the actors working together. To do this, the analyses were based on the plots and words exchanged between the designers (Figure 2). We have identified nine types of action [18]:

- listening: this operation involves taking information from a program or other participants;

- informing/sharing: this operation enables the designer to inform others and/or share their references, details of program or context;
- declaring intentions or choices/raising a question: the designer suggests and/or declares a new intention or question without trying to represent or to formalize it;
- taking action on a subject: by this action, the designer formalizes his/her intention or ideas by graphic representation;
- discussing/evaluating/questioning: this operation is reflected in the fact that an actor checks and/or discusses the proposals of another;
- validating/collective decision-making: to confirm or exclude an entire proposal related to the designed object;
- isolating: this process occurs when a group is isolated from the other group, either by choice or by the bugs, and cuts the Internet;
- coordinating/constructing the strategies of group: for this operation, the group is organized and/or sets up the meeting and / or tasks in order to work together, to validate group work strategies and/or to resolve disagreements between designers;
- intent break: this operation is involved when one actor interrupts the discussion to say something, for example, to tell a joke.

C. *Methods of Data analysis: from a generalized analysis to the specification of sequence 3.*

To approach the idea of appropriation of different tools in relation to time, the occurrence and switching, we based our studies on ideas presented above in the state of art (Section III). This is where lies the originality of our method based on complementary data: being the middlemen between two fields, that of cognitive ergonomics and semiotics, between qualitative data on the whole experiment and specific quantitative data about one particular sequence.

We relied on these three elements (time, occurrence and switches) to analyze the mode of appropriation of the system and to evaluate more precisely the stakes, the limits and the perspectives of each modality when the two tools are used.

To do this, we used a visualization tool to process our data. Called COMMON Tools, it is a web platform initiated in the framework of the ARC COMMON project and developed by LUCID of the University of Liège [19]. The tool is made available to researchers enabling them to transform data from the coding frame (in our case SketSha Replay after the coding of the data) into consolidated data then quantified and translated according to different choices of visual formalism (pie, stacked columns, time line, crossing, clouds, etc.). A range of graphics is thus proposed for the analysis of the collective design activity. It enables the formalization of quantitative data, but also to cross them (Figure 8) in relation to time, occurences, and the specificity of each actor involved in the collective design process, which corresponds totally to what we want to analyze in the context of this article.

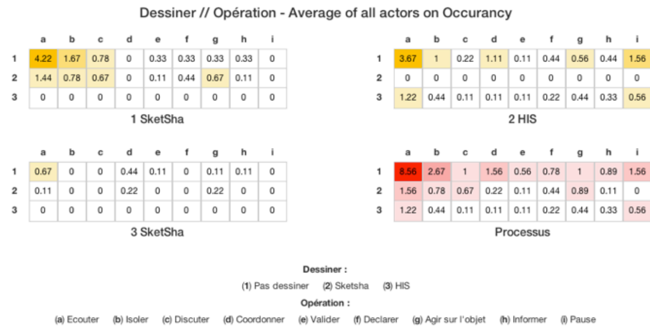


Figure 8. Example of crossing data: "Drawing" x "Operation".

V. INVOVEMENT OF THE SYSTEM IN THE PROCESS OF NEGOTIATION AND DISCUSSION-MAKING

The results emphasized in this section come from the qualitative analysis of the entire work meeting.

Here, we investigate the involvement of the two tools on the way the actors are brought to work together. It means questioning their collaboration from the point of view of leadership and the transmission of each one's ideas. To what extent can the Groupware encourage or not the collaboration or the idea of leadership? What does the synchronous use of the two tools imply the way the actors will manage a collaborative meeting? What are the risks of instrumentation? Does it mean destabilization or the fixing of ideas?

To answer these questions, we crossed two determining axes in the tooled communication activity: (1) updating through its use and (2) the contextualization. The updating through use corresponds to the way in which the potentials of the association of the two tools have been exploited. The contextualization includes the essential environmental parameters from the user's point of view.

A. *The idea of leadership in leading interactions*

The interactions between group A (with 2 student participants from Liège A1 and A2) and group B(that was mainly co-directed by the two teachers from Metz: B1 and B2) were not along the same line during all of the work meetings, between the first to the last sequence.

In fact, the hierarchical relation between the designers and their expertise relating to the tool, the training of different actors and the problem of designing experience exposed by the sponsor played an important role in the appearance of the leaders during the designing process.

In the first phase of the process (that lasted 45 minutes), groups A and B focused on understanding the order and the graphic elements at their disposal (phase of pooling of assets). Nevertheless, 2 types of leadership began to appear. The first that could be called "organizational leader", is the teacher specialized in ergonomic psychology (B2) who proposed a team work plan and who took charge of switching from one work space (HIS) to another (SketSha). It was also him who made several jokes to lighten the atmosphere or re-motivate the two groups. He was used to using these two work environments and had already manipulated them in several experiments being a researcher himself. The second that could be qualified as the 'decision - leader" is the architecture teacher (B1). Being himself in practice and being the one who is the best master in the use of SketSha (tool regularly used in the teaching context) he kept the pen almost all the time in group B. He was the one who drew the most and who made the choices ("in fact, listen… what needs to be done is..."). The actors in group A, with whom we did the self-analysis, claimed that 2 parameters should be taken into account in the B1's stand as the decision leader. The first is linked to the contextualization relating to actor B1, as much as by his knowledge of the tools as by his architectural experience. It means that a hierarchy is created by experience in the use of the device as well as in the actual design experience. When it means proposing an idea that is convergent to that of the experienced user, there should be, according to them, much more certitude and much more energy to transmit the idea. But does the active use of the device decrease the risk of an imbalance caused by the recurrence of interventions of an experienced member instead of users who have less experience?

This is where the second parameter comes into play, linked to the actualization by the use of the video-conference, by its location as well as the tone of the sound that it emits. In fact, the video-conference being completely on the right side of the screen, the actors in group A do not have direct contact with the other collaborators as was the case on SketSha (where the designers were put face to face via the vertical and centered projection of the webcam on the screen). According to them, "we lose a space of eye contact". In addition, the video-conference passes all the sounds with the same tone that makes the idea of proximity or distance lost in relation to the camera: "as if we listen to the radio … there was no question as to what they proposed… we had the impression that certain choices were imposed or that they did not listen to us...". These two phenomena due to the use of the video-conference privileged, in their opinion, the gap and the hierarchy between the actors.

The first proposal for intervention on the plan (related to the increase of reading space) came up in the 2nd phase, via some annotations on the SketSha work space. Here, also, the relationship between novices and experts was felt very strongly between Liège and Metz as much by the asides that were imposed by group B (group A did not hear anymore what was said between them) as the inter-group B or B1 kept the pen in hand during the whole phase. The actors of group A also insisted on their 'impression of rejection".

Furthermore, HIS enables one to illustrate a localized and selected point of view that depends on what is chosen and that decides to put it common. The B1 and B2 leaders, being more used to using HIS, chose their opinions, which reinforced their positions as leaders.

In the third phase, group A tried to propose a new method of work centered on the definition of the program of their future library. But this proposal was rejected by B1, followed the entire group B. A bug in the video-conference system created a feeling of being abandoned in group A. The two groups thus continued to work on the furniture to put in the library. According to them, until that point, there had not been effective collaboration since the lead in decisions had been kept by B1.

But this hierarchy was called into question in phases 4 and 5. Group A tried to show their disagreement as to the method of design that was chosen, which was supported by the organizational leader B2 who intervened in order to build consensus between the two groups: "let's go back to the objectives in the beginning… otherwise everyone will have the idea that we are not making progress". According to group A, his intervention gave back some legitimacy to their proposal, which also helped in re-balancing the statutes between the actors, especially in phase 6.

B. The transmission of ideas and the risk of destabilization

During the self-analysis, the members of group A noticed that in the first phase of the work, they had nearly forgotten the very existence of the two specific interfaces of each of the tools. The information exposed in one helped to better understand what was projected in the other. In fact, HIS enabled a life-size view of the interior spaces of the library giving a feeling of the kind of atmosphere, the quality of the present arrangements, the luminosity, etc. The view of the plan on SketSha enabled, on the other hand, to situate these views and to understand the project as a whole, with practice in interpretation and continual recognition between the view offered by HIS and that of SketSha. This observation highlights (1) the flexibility of changing from one to the other and (2) the complimentarity of the views offered by the two tools.

Gradually, as the project continues, the users referred in a casual way to each of the two tools. When an arrangement was drawn on the plan via SketSha, it was redrawn in 3D and simulated in the immersive HIS space ("this enabled us to reinforce the idea not to put little armchairs in this double-high zone… isn't there another way is to consult the book"?).

The users then went back to SketSha to increase their knowledge and realize the idea in a more global scale. HIS did not enable them to get a view of the whole, it showed a localized viewpoint selected in the space. Going back to the plan each time thus seemed necessary in this phase of the designing process.

Furthermore, by putting this experiment in parallel with their first experiment 0 using HIS and SketSha, group A explained that the way the program had been first formulated interfered a lot in their way of appropriating the tool. In fact, in the first experiment they referred to the immersive space (HIS) to decide on very precise questions whose answers demanded verification and simulation of the proposals made on the life-size plan with the help of HIS. In this case, the use of HIS was optimal, even if A1 and A2 did not master it very well. It was not the same in this experiment where the problem demanded thinking first about the program and the general objectives of the project before even rearranging the spaces. Through the association of the collaborative tools, going deeper, even if the added value of SketSha remained, the rapid zoning of the spaces and the common understanding of the whole building and organization, the added value of HIS concerned more the simulation and evaluation of the shared ideas in order to make them evolve. If, with SketSha, they had to try to better define the project and clear up certain ideas, with HIS, they could visualize in a virtual way, their hypothesis on the arrangement of spaces taking into account the scale and the double-height. In other words, they estimated that the targeted use of HIS during the entire collaboration was advantageous in the process of collective design.

Furthermore, another question was about the necessity for short chosen private discussions (and not imposed by a system bug) between the members of a team during a collaborative meeting. Did it seem to be an advantage in the eyes of the group A designers to isolate themselves to discuss together and come to an agreement before communicating their choices to the other group? In fact, during the first phase when the designers had to understand the demand and appropriate the graphic elements that they had been given, the aside, in this case, was beneficial. This moment enabled them more felt by group A, who remained passive, to agree on the feeling between them before beginning to work with group B that was geographically separated. Nevertheless, these moments must remain, in their opinion, very short to not destabilize the communication and cause a de-synchonisation in the joining of choices. This phenomenon was felt more strongly by group A, who remained passive, during phase 2 when group B, acting directly on the tool, went aside and worked without trying to interact with the others. Two types of use of the tool are to be distinguished: passive and active. From the moment of this observation, a hypothesis needed to be developed according to which a correlation is established between the two elements of time and interaction that we reformulate in the following diagram. The farther we go in the project, the more the users get involved in a collective effort of collaborative design. According to the users, a collective effort of mutual understanding is indispensable to assure the balance in the interaction and avoid destabilization of the interaction.

As for the instrumentation of the collaborative activity, the diagram shows us that the transmission of an idea can only be done through the active use of the tool once the actors are engaged in collective designing.

Consequently, it is important to reduce as far as possible the inter-team asides to succeed in transmitting one's ideas during all of the collaboration. Even if, during phase 6, this aside served the project. In fact, group A began to work on the plan and to propose a new arrangement relating to what group B was drawing on HIS. Here, the complementarity of the two tools directly served the design because two drawings emerged in parallel from two different viewpoints. One served the creation by simulation in 3D (via HIS), the other served the validation by putting the arrangements back on the 2D plan and checking that what was proposed in 3D does not contradict with the general organisation of the space in 2D (via SketSha). In the same way, group A emphasized the importance of not only synchronous, but perhaps simultaneous use of the two tools via two connected tablets: one for SketSha, the other for the space of HIS drawing.

VI. AN ADJUSTMENT PRACTICE

The results in this section focus more on the quantitative analysis of the sequence we selected for the study of the switchover from one tool to another.

Each of these sub-sequences was analyzed by using the proposed categories ("looking", "drawing" and "working together") with respect to the concept of "aspectuality." This concept allows a more accurate assessment of the issues, limitations and perspectives of each mode during the use of these two tools.

The time enables the measurement of duration of the act of looking, drawing and working together for each actor in his/her workspace. For example, depending on the relative length of the action, we distinguish two categories. The first is called "Punctual" when the designers decide to go from one tool to another. The second is related to actions that last such as when the designers discuss a problem related to the project being designed. This action is thus called "Durative".

The occurrence allows us to measure how often an action took place during the design process. In reference to semiotics, if an action is repeated (in relation to another) in a rhythmic manner and more or less orderly in a specific workspace (Sketsha or HIS), the aspectuality of this action is qualified as "Repetitive". For example, if each time an actor draws on the Sketsha tablet, the other actors look at the HIS canvas, there is repetition. If this repetition does not seem to correspond to a rule or logic, it will be qualified as being "iterative". For example, it is not systematic if an actor picks up his pen and draws to discuss an idea or to suggest a solution.

There were also cases in which the action happens only once in a specific workspace. This occurrence that denotes "single" seems significant because it can highlight the manner that a user, with regard to the tools, can appropriate his/ her work environment.

Finally, switching enabled the analysis of the data qualitatively according to the time of passage from one workspace to another (SketSha > HIS / HIS > SketSha). With reference to semiotics, aspectuality of the action is described as "Inchoate" if the action is at the beginning of a workspace. But, the action is called "Terminative" if it takes place around the end of the workspace. So, we rely on the three elements (time, occurrence and switching) to analyze the mode of appropriation of these tools.

A. Appropriation of "duplicate" and "distinctive" practices according to the time and occurrence

In this part, we distinguish duplicate practices from distinctive practices in the concept of appropriation. According to a common functionality permitted by SketSha as well as by HIS (synchronous sharing and remote graphical annotations via a tablet), actors can work together by passing from a 2D representation to an immersive representation in order to collectively design the architectural project.

The duplicate practice corresponds to the use of this common functionality between two tools. But, the distinctive practice is the use of an additional functionality. For example, the HIS also allows the use of immersive space via the 360° projector on the canvas surrounding the actors. But this immersion function is not permitted via Sketsha.

The appropriation of the use of a device combining these two systems presupposes an adjustive practice, which is halfway between duplicate practice and distinctive practice. To better understand the implications of this adjustive practice, our concern extends to the drawings, looks and words, as well as to collective operations involved in the context of architectural design activity. It must be remembered that in this experiment the actors are all invited to design a futuristic library where the need of improvement and increase of space is raised.

B. *Word exchanging, drawing and looking*

Since there is only one pen for each team, actors in the same team cannot draw at the same time on the same workspace. However, the partner who does not have a pen can "show" items on the shared tablet, he/she can "look" and comment on the projected images on immersive space and can "discuss" with all the others. As the action of "drawing" can be combined with other actions such as "looking", "showing" and "discussing", it cannot be involved except (1) in the HIS work space, (2) in the Sketsha work space. The actors can never draw simultaneously in both HIS and Sketsha workspaces. From the perspective of occurrence, the act of drawing is considered single in a workspace. But it is important to note that throughout the process, the act of drawing in Sketsha (about 10 %) is double compared to that performed on the HIS (about 5 %). The rest (85%) of the actions, which are considered as "not drawing", 1/6 of the design process in this sequence is dedicated to words and discussions between participants that are not represented graphically. Nevertheless, it becomes iterative at the end of process because when more designers advance in their choices, the percentage that is dedicated to drawing increases too.

From the point of view of time, drawing in a punctual manner corresponds to the plans' zoning. This enables actors to show zones that relate to the discussion about the project. By this action, they focus their discussions on shared graphics and make sure that all participants share the same "common ground" [20]. The act of drawing is durational when it comes to act on the subject or to discuss and evaluate potential opportunities and eventual choices for the project. By sharing this chart, they shape their discussions and synchronize cognitively the proposals of each other [21].

Therefore, drawing is done by punctual actions as well as by durative actions in both HIS and Sketsha workspaces. The punctual drawings play a demonstrative role while durative designs play an explanatory and / or argumentative role.

On the other hand, in the sequence studied, an adjustive practice specific to words, drawing and looking drew our attention. Certainly, realization of ideas happens mainly through statement and discussion between the actors because the words are meaningful, insofar as they provide elements to specify how actors contribute to the progress of the collective design. However, by comparing the action of "speaking" with "drawing", considering the time, "drawing" becomes a punctual adjustive practice during the conversations in order to clarify and explain an idea. Furthermore, aspectuality of action (durative for speaking and punctual for drawing) could be significant when combined with the activity and the space in which it operates.

Indeed, it is necessary to understand how the use of a functionality of a specific tool seems relevant or not at a specific time of collective design. The proof is the example of a designer who asked first to switch from SketSha to HIS (immersive space) because of a disagreement about the quality of light on shelves. This was then followed by a new switching when another designer requested to switch back to SketSha in order to graphically show a point that needed to be developed.

"Looking" is considered as punctual action in some cases and durative in other ones. In both work spaces, watching videoconference and looking the other participants in situ are relatively punctual actions (considering the time) but also repetitive (considering the occurrence). In HIS, we found fewer effects of going back and forth between videoconferencing and the image projected on the canvas (3D) or the one that is produced on HIS 2D tablet.

It seems that actors focus more on their annotations and graphical elements shared and produced on tablet rather than expressions of their remote partners in video conferencing or in immersive space. In occurrence, more than 3/4 of looks are directed to the workspace for the annotation in 2D. For example, "watching a videoconference" only makes participants sure about the presence of the other or about the interest of the others in conversation or the reactions of others to what has been proposed. In this case "looking at the other one who is in situ" is significant. The actors look at the others in a punctual manner (in time) but repetitive (in occurrence). " When I look at the other one, it puts my mind at rest and then I go back to my job," said one participant.

Furthermore, the action of "looking" becomes durative when one of the designers acts on the subject by using the system of SketSha for annotation. In this case, all participants look continually in the direction of the tablet. Some also look at the picture projected on the canvas.

However, when actors use only the HIS system, the one who is drawing looks rarely at the canvas (HIS 3D). He/she focuses mostly on the tablet (HIS 2D). At the same time, other users look only at the canvas on which the produced sketch is projected in 360 degrees.

"Looking at the immersive space" is involved in a punctual manner (when it comes to check punctually the validity of a choice of 2D in 3D space) and in a durative manner (when it comes to test a choice in 3D space).

In terms of occurrence, this involvement is nevertheless iterative and non-repetitive as designers look at the immersive space according to their needs and the project's progress without any apparent or pre-decided logic.

C. *Specificity of collective design*

"Evaluating", "validating", "informing" and "declaring" appear to be punctually involved in the process, while other operations (such as "listening", "discussing" and "acting on the subject") are rather durative.

Furthermore, it is important to note that the actors never tried to isolate themselves deliberately. Sometimes punctual and sometimes durative, this operation is more related to bugs caused by a network outage or disconnection of videoconferencing. However, almost all of these bugs were consistently tracked by re-questioning (via the "discussing" operation). Sometimes, they caused conflict, which, according to the users, would not have existed if the communication had been continued. Indeed, the actor interrupted by a bug is obliged to re-state what has been said before, and this sometimes causes tensions between groups.

"Isolating", "pausing" and "coordinating" operations are durative (considering the time) and iterative (considering the occurrence). They are involved here as part of the group's organization and work on several subjects for designing.

"Informing" is a punctual action whose occurrence is single in the third division in workspace (3. SketSha) while it operates iteratively in the first two divisions (1. SketSha and 2.HIS). This may be related to the project development and the mastery of problem by designers when the need for information sharing becomes less and less necessary but the action on the subject gains more importance at the end of process.

"Acting on the subject" is not only a durative operation, but also iterative because it does not follow any rule and can occur several times during the discussion.

"Validating" is punctual and repetitive because it is preceded every time by a discussion.

"Discussing" is a durative operation (by time) and iterative (by occurrence). If the operation involves a disagreement, it usually induces the request for switching from a workspace to another.

D. Appropriation relative to the time and the occurrence of the proces

Based on quantitative data from codings, we correlated, in entire process, the specificity of time and occurrence of three categories: "looking" (in Figure 9), "drawing" (in Figure 10) and "working together" (in Figure 11). These three schemes summarize the correlations for the whole process. This correlation can chart the actions and operations using both types of aspectuality; one relating to the time and the other to the occurrence.

Returning to the aspectuality of actions of each of the three sub- sequences in each workspace (see Appropriation of a "duplicate" and "distinctive" practice according to the time and occurrence) we deduced identical results.

The parallelism between these results and those put forward by charts shows that the actors appropriately duplicate practice in the same way in HIS and SketSha.

However, this parallelism is not easy concerning the distinctive practice. Indeed, we note that aspectuality is not the same from one workspace to another. If the actors refer in a punctual manner to the immersive space when they act in SketSha, they look for a long time at immersive space when switching their work to HIS.

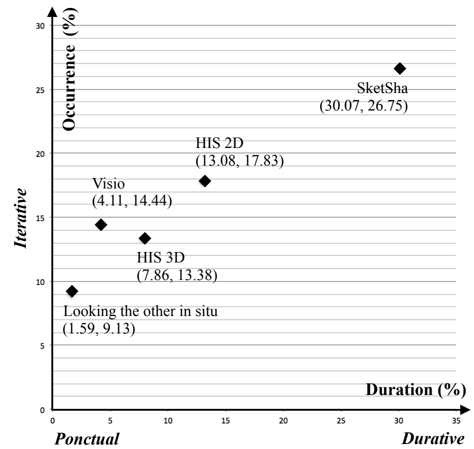


Figure 9. Correlation time/occurrence for " looking " (%).

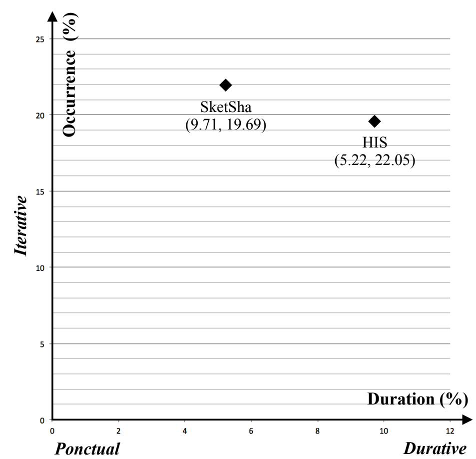
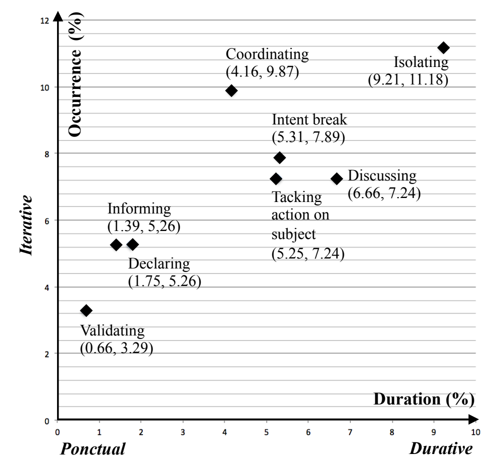


Figure 10. Correlation time/occurrence for " drawing " (%).



	Validating	Informing	Declaring	Coordi-nating	Taking action	Intent break	Discussing	Isolating	Listening
% Duration	0,66	1,39	1,75	4,16	5,25	5,3	6,66	9,21	65,61
% Occurence	3,29	5,26	5,26	9,87	7,24	7,89	7,24	11,18	42,76

Figure 11. Correlation time/occurrence for " Collective operation for design " (%).

This contrast can be explained by the degree of conformity between the functions basically provided by each tool (during their development) and uses that designers make (after combination of two tools in this experiment). The actors seem to adopt an adjustive practice (a practice between duplicate and distinctive) that seems to be in accordance with the potential of the tool and the manner it is set up by the user.

E. *Appropriation of a collective practice of switching from one tool to another*

To better understand the modes of switching from one workspace to another, we refer to the aspectuality called inchoate or terminative in this context (see Figure 3).

Qualitative analysis shows that the terminative aspect is related here to the discussion. In fact whenever there is: (1) Either a disagreement between actors about a proposal by one of them (2) Or uncertain understanding of participants about a new choice announced, designers suggest switching to another work space (from SketSha to HIS and HIS to SketSha). In this experimental context, the terminative element is imprecision and disagreement. As long as switching from HIS to SketSha is a way to check what was decided in the immersive space, actors have the opportunity to look at the same time at the canvas where annotations previously made in 3D by HIS are projected and at the tablet exposing documents and new annotations made on SketSha. So, actors can easily compare their choices for workspace. In this case, the designers are in a distinctive practice. The converse is not correct because during the switching from SketSha to HIS, the workspace for the first one disappears from the display on the tablet, and leaves the interface to the HIS workspace. The designer draws on the tablet (HIS 2D) while the other participants look at the annotation performed in the immersive space of the canvas (HIS 3D). In this second switching, designers are in a duplicate practice.

Therefore, considering the operations of "challenging" the actions performed on the object and "statement" of new proposals as terminative elements in the process of switching from one work space to another, the validation becomes an inchoate element in the process. This element marks the beginning of each switching in the use of a tool. This operation is then followed by several operations that enable the users to act on the object to be designed.

An iterative process between questioning, validation and acting on the object continues throughout the work of designers while the use of a particular tool plays a predominant role in making decisions. Indeed, even if two systems originally offer the same function for real-time and remote sharing of graphical annotations, their specificity (immersive space / non-immersive space) suggest another perspective on the object to be designed. This specificity provides a new workspace, negotiation and consensus-building between participants and allows them to test and validate their choices.

VII. Conclusion

Contribution. Our research concerned the modes of appropriation of an innovative collaborative platform, to instrument distant and synchronic design by associating two tools, which support artifact annotation in real time.

This work allowed us to develop an analytical method that uses concepts related to semiotics in order to observe systemically the collective activity of design using various tools at the same time. In fact, through our data analysis and by using this method at the border of the fields of cognitive ergonomics and semiotics, we could clearly identify the use of 2D, the use of 3D and switching from one to another. In other words, what makes an actor switch from one to another? The observation of this practice that is at once "duplicate" and "distinctive" showed that look, drawing, and word (representing "working together") play an important role.

It is obviously possible to draw in a tool and look simultaneously at another workspace, and this was observed during the use of SketSha (2D plans on tablet produced parallel to the interior image of library, which was projected in the immersive space. In this case, it was not a switching from one tool to the other but an oversizing of the workspace. The activity was not just in 2D or 3D, but it was oversized to offer two different perspectives simultaneously for a single area of the designed object. Even when actors worked in space dedicated to SketSha, they occasionally referred to the immersive space. However, in the context of use of the HIS device, the interface of HIS 2D appearing on the tablet involves systematically the disappearance of the SketSha workspace. A definite switching from one activity on a tool to a new activity on another tool is marked.

Moreover, aspectuality related to switching of certain collective operations shows the effectiveness of the combination of two tools in order to validate the collective choice in the collective and remote design of a project. In both cases of switching (1) from SketSha to the HIS and (2) from HIS to SketSha, appropriation of a tool's specific functionality allows designers to better understand the ideas expressed, to build a common ground and to move forward together in a preliminary design phase. Nevertheless, the recurrent problem of bugs and sound dropping during the videoconferencing due to network disconnection did not help to build awareness among participants. This even caused some conflict between them. Both findings highlight the notions of completion and accomplishment throughout a permanent evaluation of ideas in the process. If all the operations that we have emphasized are essential in these early stages, it would still be considered a privileged place for punctual operations such as "informing", "declaring" or "validating together", which require good functionality of the tool.

Limitations. Focusing on the modes of simultaneous appropriation of these two tools for collaborative design, this research is certainly not intended to be generalizable to other cases of tool and device combination. Nevertheless, the method implemented for processing and analysis of this type of combination is still interesting because it combines quantitative and qualitative data in a systematic, repeatable and disciplined approach. To further this approach and prove its validity, it is necessary to confront other contexts of using combined tools by exploiting the concepts from the field of semiotics.

In addition, semi-structured interviews were conducted as part of this experiment, but these data were only used partially in our analysis.

The in-depth processing of designers' feedback will enable the issue of aspectuality to be addressed in greater detail from the users' perspective by reference to how they describe their experiences of appropriation of combined tools.

Prospects. We plan to apply our approach (1) on one hand in longitudinal observations to analyze the evolution of this appropriation process in time and (2) on the other hand, to observe new collective activities such as participative production of a same artwork from distance.

REFERENCES

[1] S. Ben Rajeb, S. Shirkhodaei, and P. Leclercq, "Time, Occurence and Switching, Appropriation of two tools in collaborative design Point of view of aspectualization", proceedings of the Seventh International Conference on Advanced Collaborative Networks, Systems and Applications, COLLA'17, Nice, France, 2017.

[2] T. Dorta, Y. Kalay, A. Lesage, and P. Pérez, "Design conversations in the interconnected HIS", in Carrara G., Fioravanti A., Trento A. (eds.), Connecting Brains Shaping the World: Collaborative Design Spaces, EuropIA 13, Rome, pp. 83-94, 2011.

[3] http://www.hybridlab.umontreal.ca/his.php?l=fr, last access in April 2017.

[4] S. Ben Rajeb and P. Leclercq, "Using Spatial Augmented Reality in Synchronous Collaborative Design", Yuhua Luo éd., In Lecture Notes in Computer Sciences, Springer, vol 8091, pp. 1-10, 2013.

[5] http://www.sketsha.be/F11_Innovation.html, last access in May 2017.

[6] J. G. Breslin, "Social Semantic Information Spaces", in Semantic Digital Libraries, pp. 55-70, 2008.

[7] M. Lourel, S. Abdellaoui, S. Chevaleyre, M. Paltrier, and K. Gana. "Relationships between psychological job demands, job control and burnout among Firefighters", in North American Journal of Psychology, 10(3), 489-496, 2008.

[8] F. Darses, M. Wolff. "How do designers represent to themselves the users' needs?", in Applied Ergonomics, 37 (6), pp. 757-764, 2006.

[9] J. F. Boujut, P. Laureillard, "A co-operation framework for product process integration in engineering design", in Design Studies, 23 (5), pp. 497-513, 2002.

[10] P. Carlile, "Transferring, translating and transforming: an integrative framework for managing knowledge across boundaries", in Organization Science, 15, pp. 555-568, 2004.

[11] J. M. Carroll, D. C. Neale, P. L. Isenhour, M. B. Rosson, and D. S. McCrickard, "Notification and awareness: synchronizing task-oriented collaborative activity", in International Journal Of Human-Computer Studies, 58, pp. 605-632, 2003.

[12] A. J. Greimas and J. Courtés, Sémiotique. Dictionnaire raisonné de la théorie du langage, (translated by "Semiotic. Dictionary of the Language Theory"), Paris, Hachette, 1993.

[13] J. Fontanille, Le discours aspectualisé, PULIM, Limoges, 1991.

[14] J. Holt, Etude d'aspect. (translated by "Aspect studies"), Acte Julandicia. 15: 2 Copenhague, 1942.

[15] J. Dubois, M. Giacomo, L. Guespin, Ch. Marcellesi, J.B., Marcellesi, and J.P. Mével, Grand Dictionnaire de linguistique et des sciences du langage, (translated by "Dictionary of Linguistics and Language Science"), Paris, Larousse, 2007.

[16] D. Bertrand, Précis de sémiotique littéraire (translated by "Literary semiotics"), Paris, HER, 2000.

[17] S. Ben Rajeb, "Modélisation de la collaboration distante dans les pratiques de conception architecturale: Caractérisation des opérations cognitives en conception collaborative instrumentée". PhD thesis in architecture. Laboratoire Modélisations pour l'Assistance à l'Activité Cognitive de la Conception (MAP-MAACC de UMR MAP CNRS 3495), Université Paris-Est et Ecole Nationale d'Architecture de Paris La Villette, Paris, France, 2012.

[18] S. Ben Rajeb, "Conception collaborative distante: étude architecturologique pour la caractérisation des opérations cognitives" (translated by "Remote collaborative design: architecturology analysis to specify cognitives design operations"), V. Angenot, S. Safin, M.G. Dondero & P. Leclercq éd., In Interfaces numériques: Collaborer à distance, enjeux et impacts des interfaces numériques dans les pratiques collaboratives synchrone, Lavoisier, 2 – n°3, pp. 509-530, 2013.

[19] S. Ben Rajeb and P. Leclercq, "Instrumented analysis method for collaboration activities", proceedings of the Fifth International Conference on Advanced Collaborative Networks, Systems and Applications, COLLA'15, San Julian, Malta, 2015.

[20] J. Beers Pieter, H. P. Boshuizen, P. A. Kirschner, and W. H. Gijselaers, "The analysis of negotiation of common ground in CSCL ". Learning and instruction, vol. 17, n°4, p. 427-435, 2007.

[21] F. Darses, A. Mayeur, C. Elsen, C., and P. Leclercq, "Is there anything to expect from 3D views in sketching support tools?". Gero, J., Goel, A. (eds.) in Design Computing and Cognition 2008, proceedings of the Third International Conference on Design Computing and Cognition, Atlanta, USA, pp. 283–302. Springer, 2008.

User-guided Graph Exploration: A Framework for Algorithmic Complexity Reduction in Large Data Sets

Tim Grube*, Florian Volk†, Max Mühlhäuser*, Suhas Bhairav‡, Vinay Sachidananda‡, Yuval Elovici‡
*Telecooperation Group, TU Darmstadt, Germany
†House of IT, Germany
‡iTrust, SUTD, Singapore
Email: {grube, max}@tk.tu-darmstadt.de, volk@house-of-it.eu, {suhas_setikere, sachidananda, yuval_elovici}@sutd.edu.sg

***Abstract*—Human exploration of large data sets becomes increasingly difficult with growing amounts of data. For this purpose, such data sets are often visualized as large graphs, depicting information and interrelations as interconnected vertices. A visual representation of such large graphs (for example, social networks, collaboration analyses or biological data sets) has to find a trade-off between showing details in a magnified—or zoomed-in—view and the overall graph structure. Showing these two aspects at the same time results in a visual overload that is largely inaccessible to human users. In this article, we augment previous work and present a new approach to address this overload by combining and extending graph-theoretic properties with community detection algorithms. Our non-destructive approach to reducing visual complexity while retaining core properties of the given graph is user-guided and semi-automated. The results yielded by applying our approach to large real-world network data sets reveal a massive reduction of displayed vertices and connections while keeping essential graph structures intact.**

Keywords*—*Complexity reduction; graph visualisation; big data exploration; graph metrics; community detection.

I. Introduction

Processing data sets is becoming increasingly easier. With the rise of *big data* and powerful computing devices, the collection and processing of large data sets has become a common thing in both research and industry. This article extends beyond our previous work on reducing visual complexity in graphs [1] with the introducing of our framework and an extended evaluation.

Many fields profit from the capability to reveal new insights and connections that can only be detected by analyzing large amounts of collected data.

However, Moore's Law [2] does not apply to the ability of human users to understand and explore such big data sets. Making large data sets accessible to human users is difficult and becomes increasingly more difficult with ever-growing data sets. Human-centric data analysis techniques usually employ visualization of the data sets. Visualizing real-world data sets as graphs quickly results in an inaccessible chaos of large amounts of interconnected vertices due to the large amount of information to be shown. See Figure 1 for a visualization of a part of the Facebook social network [3]. In order to comprehend relations within such a graph, a user needs to magnify the visualization a lot. This magnification implies a loss of overview, so that a user might be able to understand specific relations within a part of the graph, but loses track of the overall structure.

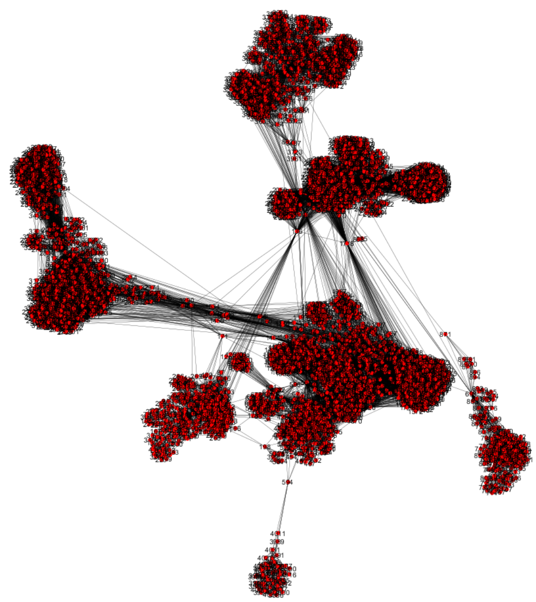


Fig. 1: User relations in a social network (data taken from Facebook).

This issue of simultaneous visualization of both details and overall structure complicates the process of exploring data sets when the intended outcome is unknown. This is a common scenario for exploratory data analysis. As such explorative scenarios often call for human creativity as the aim is to find new insights, connections and cross-references that are not yet algorithmically formalized. Pre-defined and well-specified analyses can be processed automatically without human involvement, and thus, without the need for visual representation. However, exploratory data analysis usually does not allow for automated pre-filtering of data sets to obtain a human-centric, *decluttered* view on the data and the *information* to be discovered.

The framework and methodology we present in this paper combine and extend graph-theoretic properties with community detection. The assumption is that concealing less important vertices (and their connections) from a graph leads to a compressed and easier-to-understand visual representation that still contains the information intended to be discovered. The framework is designed to assist a user-guided iterative process of data reduction. For every iteration step, the user first selects a vertex as template, thereby defining a *selection criteria value* (SCV). All other vertices are filtered depending on their SCV in comparison with the selected SCV. Vertices of lower importance are concealed while ones with equal or higher importance are retained. This step can be combined with community detection, which reduces clusters of similar vertices to a single vertex, retaining the structure of the graph, but not the data density. This semi-automated, non-destructive approach to reduce visual complexity is assumed to identify core insights to be yielded from the data while side-information is hidden.

The proposed framework and methodology aim at improving human-lead analysis and understanding of huge data sets; efficiency and scalability aspects in automated processing of such data sets are not in the focus of this research.

Our results indicate a massive reduction of displayed vertices and links between them in every iteration of our proposed approach. Therefore, only one to three iterations effectively reduce graphs as the one shown in Figure 1 to a representation that is easy to comprehend for human users (see Figure 3 for the results of just three iterations).

The article is structured as follows: in Section II, we discuss relevant related work on complexity reduction in graphs; in Section III, we discuss fundamental methodological background information to our contribution by introducing the graph-theoretic metrics and community detection mechanisms incorporated in our framework. For every method, we discuss why and how we assume it is beneficial to our goal of removing unnecessary data. Section IV, the general approach to apply community detection and SCVs to iteratively filter vertices is presented. Our framework itself is presented in Section V, discussing both the internal structure and the visual user interface. Our assumptions and framework are experimentally evaluated in Section VI, in which we describe the simulation setup and discuss the obtained results in detail for various data sets. We conclude with a summary and future research directions in Section VII.

II. State of the Art

In this section, we discuss related work with respect to the field of complexity reduction in graphs. Most of these contributions are related to the field of human computer interaction (HCI) and aim on making complex datasets accessible to humans.

Kimelman et al. [4] proposed techniques like ghosting, hiding and grouping of edges. The vertices and edges of the graph were removed based on various techniques like weights of edges, labels of vertices, etc., but these techniques were concerned with dynamic graphs. Differently, Holten et al. [5] reduced only the visual cluttering of edges by bundling them.

Fisheye techniques [6] [7] tend to concentrate only on the interesting regions of the user. The zooming feature in such techniques is only responsible for making a very small region of a graph appear larger. They do not remove any vertices or edges. So, the overall graph content remains the same. Fisheye views that retain structure are introduced by Furnas [8]. Abello et al. [9] introduced hierarchical clustering and depiction of a tree-map in addition to a compound fisheye view technique but never concentrated on reducing the overall size of a network.

Various approaches towards creating communities in large graphs are presented in [10] [11] [12]. These techniques provide a significant level of understanding of the kind of vertices and their properties in large networks but never used the same to reduce the overall content in a large network and provide a simpler view.

Sundararajan et al. [13] introduced rectangular partitioning and Voronoi partitioning techniques. The former involved partitioning the area of display into four quadrants while the latter involved the partition area being closer to the concerned vertex. This only reduces the distortion in the graphs.

Batagelj et al. [14] took a mathematical approach through the usage of matrix. Large graphs were reduced to *k-cores*. Later, the graphs are represented as an adjacency matrix or a contextual matrix based on their size. But when the graphs grows really large, managing the matrix becomes a enormous task.

All before mentioned techniques reduce complexity based on a global perspective, disregarding the current interests of a user. Thus, these techniques may maintain and highlight information that is not relevant the user's current situation while rejecting important information from the user's slant.

III. Methods

In this section we introduce our terminology and the various graph-theoretic metrics we apply to measure the importance of vertices within a graph. The reason hereby is to distinguish important vertices that are relevant for the human observer from those vertices that can be removed from the visualization. Less important vertices are considered to be non-essential for the visualization and, as such, can be removed from the visualization without altering the general information the graph explorer is looking for.

A. Terminology: Graph-based Data Representation

Most data sets can be represented by a Graph $G = (V, E)$ that is formed by a set of *vertices* $v_i \in V$, which represents the pieces of data. Relations and *connections* between these pieces of data are represented by *edges* $E \subseteq V \times V$, i.e., connections are represented by pairs of vertices (v_x, v_y). Considering our exemplary application scenario of a social network, each vertex represents a user and edges represent the connections between users. Another example is the model of (research) citations as citation graph where publications are represented as vertices, and a citation is represented by a directed edge.

A graph G can be *directed* or *undirected*. In a directed graph, an edge e_k can exist between a vertex v_i and v_j while the other direction (e_l: v_j to v_i) may or may not exist, independent of the existence of e_k. In an undirected graph, the existence of before-mentioned edge e_k implies the existence of e_l. A social network like Facebook applies undirected connections, thus, if user Alice is connected to user Bob, Bob is also connected to Alice (Alice and Bob are "friends"); in contrast to that, Twitter applies directed connections. Thus, Alice may be connected to Bob (Alice "follows" Bob), but Bob may not be connected to Alice.

The number of connections of a vertex v_i is noted as the *degree* d_{v_i} of v_i. In case of a directed graph, the degree of a vertex has to be specified for outgoing connections: the *out-degree* and for incoming connections: the *in-degree*.

A connection between any two vertices is called a path p. A path between adjacent vertices has the length 1; yet, a path may include intermediate vertices to connect them. A path p is represented by an sequence of edges $p = (e_0, e_1, ..., e_n)$. The *shortest path* between two vertices is the path with least edges.

B. Selection Criteria Values

Vertex selection is one of the fundamental steps in reducing the complexity of a graph, i.e., it is the selection of an appropriate subgraph. Selecting a set of vertices based on specific parameters helps in creating a subgraph that retains the inherent properties of the graph and reflects the user's interests. The vertices are identified based on *selection criteria values*. An SCV acts as the basis of vertex selection in our framework. For a vertex to be selected as part of a graph, its SCV must be greater than the SCV of the user's interest. Here, we express the user's interest by selecting a (start) vertex whose SCV is compared with every other vertex in the graph.

The SCVs include graph-theoretic properties and centrality measures but are not limited to just these presented measures. We focus this paper on the usage of importance, connectivity, and distance measures as these appeared most promising in our literature research. The importance measures are PageRank and Betweenness Centrality; the connectivity measures are Clustering Coefficient and Degree Centrality; the distance measure is Closeness Centrality. In the following sections, we discuss every SCV to understand their significance to our idea of reducing visual complexity.

1) Closeness Centrality: Closeness centrality determines the closeness of vertices in a network, i.e., it determines the distance of vertices in a graph. The vertices that have high closeness centrality values are considered to be closer to other vertices in the graph. For example, in a network where information flows from one vertex to another, transmission of information takes place quickly due to their high closeness centrality values.

The rationale for using closeness centrality is as follows: a set of vertices that are close to each other may form a center of information within a graph. Close neighbors are thereby thought to attribute to the general information that a human user is looking for. Vertices that are far away from this information center only contain auxiliary information and can be omitted during graph exploration. Closeness is generally attributed to the quickness in flow of information in the network. The quicker the information arrives at or departs from a vertex, the closer is the destination or the source vertex respectively [15].

According to Freeman [16], closeness centrality of a vertex is defined as:

> *"The sum of graph-theoretic distances from all other vertices, where the distance from a vertex to another is defined as the length of the shortest path from one to the other."*

For example, in a social network scenario, this could mean when an information is shared between two individuals, the chances of such information propagating in the network is higher with those who are immediate friends or neighbors. For a given vertex p_i, closeness centrality [17] is calculated using (1), where $|V|$ denotes the total number of vertices in the graph, i and k are integers where $i \neq k$, $d(p_i, p_K)$ denotes the number of edges in the shortest path between p_i and p_k.

$$C_C(p_i) = \frac{|V| - 1}{\sum_{k=1}^{V} d(p_i, p_k)} \quad (1)$$

2) Betweenness Centrality: Betweenness centrality depicts influence or powerfulness of a vertex in a network. The vertices that have high betweenness centrality values are *highly influential* as they act as bridges or shortcuts for interactions between several pairs of vertices or even information centers. Vertices with low betweenness centrality are considered to be less influential than others and can be removed from a graph without changing the structure of its information flow.

According to Freeman [16], the betweenness centrality of a vertex is defined as:

> *"The share of times that a vertex i needs a vertex k (whose centrality is being measured) in order to reach a vertex j via the shortest path."*

For example, in a social network scenario, where individuals form friendships with individuals who in turn have a large group of friends tend to have an influential clout due to their connections with such individuals.

Betweenness centrality [17] can be calculated using (2), where g_{jk} denotes the total number of shortest paths between p_i and p_k, and $g_{jk}(p_i)$ denotes the number of shortest paths containing p_i.

$$C_B(p_i) = \sum_{i \neq j \neq k \in V} \frac{g_{jk}(p_i)}{g_{jk}} \quad (2)$$

3) PageRank: PageRank [18] provides a rank to every vertex in a graph. It extends the idea of computing citations to a web page by ensuring that all rank values are normalized based on the total number of links on a page instead of treating all web pages equally.

A transition matrix is created based on the transfer of importance from one to another. Initially, we apply a uniform distribution based on the initial grade structure. Then, depending upon the incoming connections we re-calculate the PageRank value [19].

A high number of incoming links is thought to indicate high importance of the respective vertex. Transitive importance is given by PageRank as those who have incoming links from vertices with a high PageRank inherit this importance. As such, vertices with low PageRank values are not closely connected to those vertices, which contain the core information in a graph. By only keeping the important vertices within a graph, this core information is retained even if all other vertices are hidden from the human user.

The most prominent example for PageRank application is the world wide web network: a web page that is referenced in many other web pages will have a high PageRank value. In a social network scenario, individuals with higher PageRank are generally highly powerful due to a lot of connections whose betweenness centrality value is also high. The concept of PageRank is closely related to the betweenness centrality as it measures the importance by the number of vertices that are (transitively) pointing towards another one; the transitive propagation of importance is, however, unique to the PageRank.

4) Degree Centrality: Degree centrality portrays the level of connectivity of a vertex in a network. The vertices that have high degree centrality values are more influential or important. In difference to PageRank, degree centrality is not inherited from connected vertices, but only measures a degree of connectedness. Anyhow, similar to PageRank, highly connected vertices are considered to be of higher importance to retaining the information in a graph than vertices with only few connections. The latter ones are hidden from the user by our framework.

The degree centrality is computed as presented in (3) and given by the number of adjacent neighbors.

$$C_D(p_i) = deg(p_i) \tag{3}$$

For example, in a social network, a vertex with a high number of connections is powerful and highly visible when compared to others [11].

5) Clustering Coefficient: As stated before for the centrality metrics, we assume the existence of information centers within graphs. Our framework is tailored towards identifying and retaining these information centers while removing as many less important vertices as possible from the graph.

The clustering coefficient gives an indication of clusters in a network. The higher the clustering coefficient, the closer a vertex is to a cluster, i.e., vertices with a high clustering coefficient are parts of clusters. Vertices with a low clustering coefficient are considered outliers with respect to information centers and, therefore, can be removed from the graph without a change to the information a user is looking for.

According to Zafarani et al. [20], the clustering coefficient can be defined as shown in (4).

$$CC = \frac{3 \times (Number\ of\ triangles)}{Number\ of\ Connected\ Triples\ of\ Vertices} \tag{4}$$

For example, in a social network, vertices with high clustering coefficients have a desire to form close bonds or friendships with their neighbors [21].

C. Community Detection Algorithms

Community detection allows to group vertices that share similar properties. The properties vary depending on the type of the used *community detection algorithm* (CDA). For example, Louvain algorithm optimizes communities with respect to maximizing their modularity while direct K-means community detection optimizes on the spread of communities.

We use the found communities to reduce complexity in larger steps, being able to remove larger portions of the graph in the first few complexity reduction cycles. CDAs form one of the core evaluation criteria in our implementation. The four algorithms used in our framework are Hierarchical Clustering, Original Louvain, Louvain with Multilevel Refinement, and Direct Clustering. In the following sections, we will discuss the CDAs that help us in understanding the way in which various clusters are formed.

1) Hierarchical Clustering: The hierarchical clustering method by Jain and Dubes [22] involves obtaining a series of partitions that are nested by transforming a proximity matrix. The proximity matrix represents the distance of the individual vertices.

Based on this matrix, vertices are merged in an iterative process. First, the two closest, trivial clusters of one vertex are merged; after that, the next two closest clusters are merged. As the algorithm is applied, a hierarchy of clusters is established. Those clusters are usually visualized using a dendrogram, as shown in Figure 2.

We utilize the complete-link clustering method in an agglomerative (bottom up) setting in our implementation. One

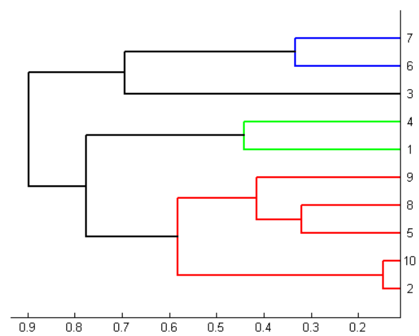


Fig. 2: Hierarchical Agglomerative Clustering visualized.

of the main advantages of this method is the avoidance of chaining-effects. Also, the clusters are balanced and smaller when applying this method.

2) Original Louvain: The original Louvain algorithm by Blondel et al. [23] detects communities in two steps that are repeated:

1) establishing communities by locally grouping vertices such that the graph *modularity* [24] is maximized. For the modularity maximization, vertices are moved between communities [25].
2) replacing the communities with single vertices to reduce the complexity of the graph.

These steps are repeated until no improvement of modularity is achievable and a hierarchy of communities is established.

3) Louvain with Multilevel Refinement: Louvain with Multilevel Refinement algorithm was formulated and conceived by Rotta et al. [26]. The algorithm is similar to the original Louvain community detection, but the adds a *refinement* phase at the end to the community detection algorithm. In this refinement phase, the achieved clustering hierarchy of the original Louvain method is visited in reverse order, i.e., from the coarsest communities to the trivial communities.

This multilevel approach provides communities exhibiting better modularity values at the cost of higher computation effort.

4) Direct K-means: The direct K-means algorithm by Alsabti et al. [27] establishes k communities. For that, the following steps are repeated until the algorithm converges to its solution:

1) k vertices are selected at random and form the centers of the k communities.
2) the remaining vertices are assigned to their closest community.
3) for each community, it is evaluated, which vertex is the actual center of the community by computing the mean position.
4) the vertices that are closest to the mean position of their respective community are selected as the center of their community.
5) repeat from step 2)

IV. Complexity Reduction Approach

In this section, we present our iterative and human-centric approach to reduce the data complexity in graphs using SCVs and CDAs.

A. Subjectivity of Visual Complexity

Visual complexity is a subjective measure. Every user has a different understanding of visual complexity of a data set, depending on previous experience, expert knowledge, and familiarity with the respective subject. Hence, the challenge of visual complexity reduction cannot be solved with a "one for all" approach.

To overcome this subjectivity challenge, we provide a user-interactive round-based complexity reduction method that allows to reduce the visual complexity in succeeding iterations until the user is able to understand the data set well enough to collect the desired information.

B. Start Vertex Selection

Selecting a start vertex is highly dependent on the analysis goal of the user. Thus, prior knowledge beyond the scope of this paper is required. However, this knowledge is available to the user. That is why the proposed approach employs an interactive start vertex selection, which (i) suits the analysis goal as well as (ii) solves the issue of aquiring appropriate prior knowledge. The interactive start vertex selection may be implemented using a graphical repesentation of the overall graph, in which a human user simply points and clicks on the desired start vertex. Otherwise, explicitely naming or describing the vertex by a set of properties or by a name is possible as well.

The evaluation in section VI is non-interactive. Therefore, a random start vertex is selected in every iteration. As the evaluation is to analyse the effectiveness of visual complexity reduction, selecting a "random" analysis goal seems reasonable.

C. Visual Complexity Reduction

Our method to reduce the visual complexity is a two-step process, which also allows to parallelize the computation to reduce computation time. We give an algorithmic description in Algorithm 1. In the first step, we apply a community detection algorithm to group similar information. According to lines 3-6 of Algorithm 1, we compute the disjoint communities and proceed with each of them individually. At this point, we

```
function ReduceVisualComplexity(g, r, s, id, type_g, c_alg)
    g: Input graph
    r: Number of iterations
    s: Selection criteria
    id: Start vertex (picked by user or at random)
    type_g: Graph Type (directed or undirected)
    c_alg: Community detection algorithm
    g_reduced: Output graph with reduced complexity

    g_reduced ← new Graph();
    for i ← 1 to r do
                                        ▷ loc: List of Communities
        loc ← getCommunities(id, g, algo, g_type)
        for j ← 1 to sizeOf(loc) do
            community ← loc.get(j)
            for k ← 1 to sizeOf(community) do
                v_k ← community[k]
                if getSCV(v_k, s) > getSCV(id, s) then
                                        ▷ compute shortest path
                    sp ← getSP(v_k, id)
                    addToReducedGraph(sp, g_reduced)
                            ▷ break if applied w/ community detection
                    if loc > 0 then
                        break loop
                    end if
                end if
            end for
        end for
        i++
    end for
    return g_reduced
end function
```

Algorithm 1: Iterative Reduction of Visual Complexity.

can easily parallelize the computation, having the communities being handled concurrently.

In the second step, we have to differentiate whether community detection was applied in the first place or not. If community detection is not applied, we calculate the SCV of each vertex and compare it with the SCV of the vertex representing the user's interest. When the SCV of the inspected vertex is higher than a threshold, which is in our case the SCV of the user's interest, the vertex is retained in the reduced graph. To preserve the relation and connection of the user's interest and inspected vertex, we retain not only the vertex but also the vertices on the shortest path. In our social network scenario, these vertices represent the chain of friends between two subjects and, thus, may be relevant to the user.

If community detection was applied in the first place, we only retain representatives of each community and hide the remaining vertices of a community. This way, we can easily achieve a massive complexity reduction in only a few iterations. Yet, the user can select the representative of a community as next interest, restarting the process from this new perspective and gain new insights into the data set.

These two steps are then repeated until the user is able to sufficiently comprehend the graph. Our algorithmic description uses a predefined number of iterations, yet an application would delegate this decision to the user, who might request further iterations.

In Figure 3, we visualize the iterative complexity reduction process. On the left-hand side, the original graph, which fed into the visual complexity reduction mechanism, is visualized. In three steps, visualized towards the right-hand side, we reduce the visual complexity of the data set considering a randomly chosen vertex to reflect an otherwise user-selected interest. The user interest is marked by the green vertex with the ID 12 in Figure 3. The original data set, labeled "Start", is the Facebook data set already presented in the introduction in Figure 1.

D. Optional Application of CDAs

The application of CDAs is optional in our framework and approach. If the user assumes the presence of "distinct" communities or clusters, the application of CDAs helps to preserve this structure.

Without application of CDAs, the SCVs focus on absolute values that they compute and, therefore, they select the absolut top vertices according to their graph metric. In combination with CDAs, the representative vertices that are retained in each iteration are selected from the different communities. Thus, a higher diversity is achieved.

However, without inherent community structure in the graph, the application of CDAs may be a) misleading and b) a waste of computing resources.

V. Framework

In this section, we describe design decisions and components that are considered while developing the complexity reduction framework. The framework is released to the research community on a public GitHub repository [28].

A. Internal Structure

Our framework for complexity reduction is based on the graph analysis framework *GraphStream* [29], a framework by researchers of the University of Le Havre and members of the RI2C research team. GraphStream is extended by the CDAs and the SCVs and a graphical user interface to enable human users to explore complex data sets visually.

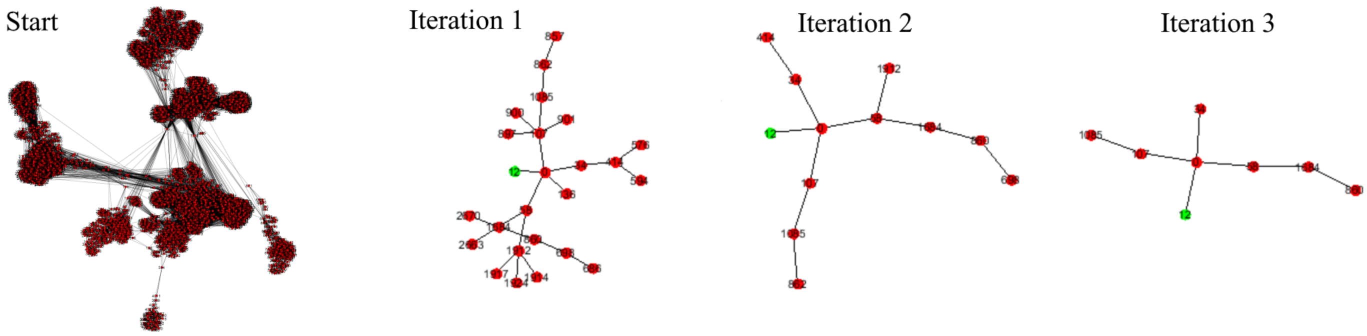


Fig. 3: Complexity Reduction of Graph in Fig. 1 by applying Algorithm 1 (first three iterations shown).

1) GraphStream: GraphStream is a Java-based library for modeling and analyzing dynamic graphs. For that, GraphStream provides a graph model that enables vertices and connections between them to appear and to disappear over time. Static data sets and graphs can be modeled by ignoring the dynamic capabilities of GraphStream.

GraphStream provides capabilities to visualize—and therefore, to layout—graph-based data sets and to calculate graph-theoretic measures to analyze and understand graph-based data sets; moreover, by using the dynamic capabilities of GraphStream, we are able to alter the data sets to reduce their visual complexity.

a) Data Handling: Users can load their specific data set into the application. Today, the dataset should be structured according to the `dgs`-format of GraphStream; in this format, a graph and its dynamic changes are described according to time steps (or clock ticks) and a number of changes (events) within in these intervals—a static graph is described by skipping the time step declaration such that every change (event) is performed at the beginning. The format is as follows:

```
DGS004 .........................◁ file format & version
<name> <#time steps> <#events>
st 0 ................................◁ time step w/ id 0
an n1 ..................◁ add vertex n1 w/o parameter
an n2 x=1, y=3  ◁ add vertex n2 w/ parameters x, y
cn n2 x=2, y=2 .....◁ change parameters x, y of n2
dn n2 ................................◁ delete vertex n2
ae n1 n2 .............◁ add connection between n1 n2
ae n1 > n2 ..........◁ add connection from n1 to n2
ae n1 < n2 ..........◁ add connection from n2 to n1
ce e1 w=10 ....◁ change parameter of connection e1
de e1 ............................◁ delete connection e1
```

Providing a continuous flow of these events, for example, using a network interface, dynamic changes of the data base can visualized.

To enable the subsequent evaluation in this article, a reader for simple adjacency lists (i.e., a list where each entry is a tuple “n1 n2” representing a connection between vertices n1 and n2) is implemented.

More readers for different file formats can be added easily, which is a necessity when considering the usage of our framework in vastly different application scenarios. For that, the user has to provide a file parser that generates the vertices and connections between vertices according to their very own specifications and semantic.

b) Layout and Visualization: The standard layout algorithm in GraphStream is force-based, i.e., the layout algorithm in GraphStream applies repelling forces to each vertex and contracting forces to each connection. These forces ensure that vertices are placed with distance to each other and, yet, connected vertices are grouped.

Force-based visualization already supports human users in understanding large and complex data sets by arranging the graph according to its connection-structure and its clusters, for example, the initial graph in Figure 3 reveals seven groups of vertices. By this grouping, the distance between groups and the placement and concentration of connections, a human user is able to get first insights by taking his application knowledge into account.

2) Control Flow: The internal control flow in the framework is depicted in Figure 4. First, the user loads a data set into the framework. Using the graphical user interface, the user can then configure the complexity reduction by choosing a selection criteria value and an (optional) community detection algorithm. With the selection of their interest, the user passes the control to the complexity reduction module, which is composed of the community detection (using the CDAs) and computation of the SCVs.

In the complexity reduction module, the optional communities are established and the representatives for each community are computed; after that, the vertices are removed accordingly. This step is then repeated until either of the following termination conditions is fulfilled:

- *Iterations done:* after n iterations, the complexity reduction step is completed and may be restarted by the user.
- *No change:* if two subsequent graph representations are equal, i.e., computing communities and selection criteria values does not enable the removal of additional vertices, the complexity reduction is completed. By selecting anot-

her vertex as interest, the user may restart the complexity reduction.

After completing the complexity reduction, the control is passed back to the user interface where the reduced data set is visualized to the user. From here, the user can refine their interest and select another start vertex for another complexity reduction.

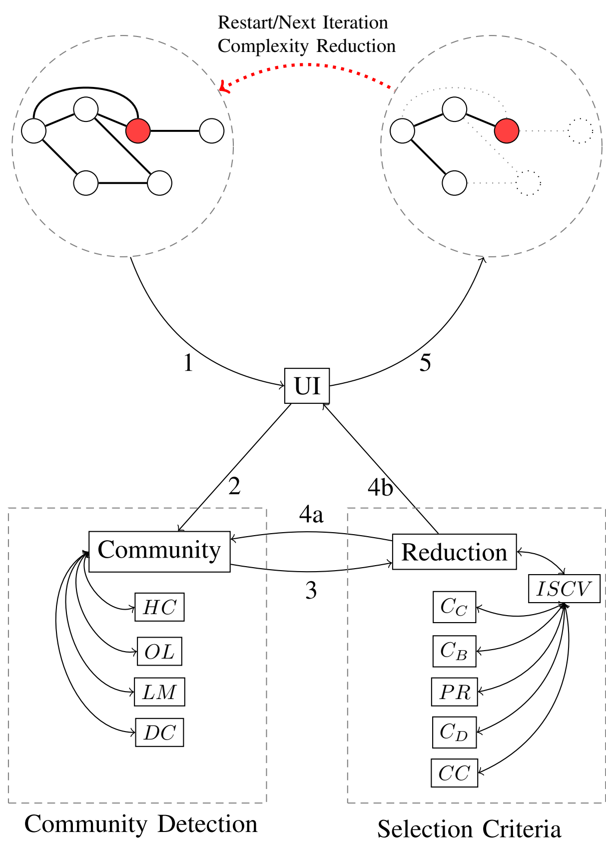


Fig. 4: Control Flow.

3) CDAs and SCVs as Modules: The CDAs and SCVs are the core components of our framework. By their ranking of vertices, CDAs and SCVs do also define, which properties of vertices are distinguishing the vertices to preserve from the ones to remove.

The actual properties defining the importance of vertices are highly depending on the application scenario and use case, i.e., the type of information the user is looking for. For example:

> *"Reducing complexity in a social network when searching 'influential people' will look for dense communities and the highly connected people in these communities."*

To account for this variability of required properties, the implementation of CDAs and SCVs follows a modular structure that allows easy extension by researchers or users of the framework.

The CDAs are encapsulated in their own class, the actual implementation of the community detection algorithms is based on the `S-Space` [30] repository of Jurgens and Stevens from UCLA.

The SCVs are hidden behind the interface `ISelectionCriteria` to provide their functionality without enforcing a stronger coupling of different modules with the core framework.

B. User Interface

The focus of the user interface is the visualization of the graph. From this perspective, the user is able to explore their data according to their application scenario. Figure 5 visualizes the framework from a user's perspective.

From the initial view, the user has the ability to load a data set provided in the aforementioned `dgs`-format—or to generate a synthetic network, for example a *Barabàsi-Albert* network.

After visualizing the initial "root" graph, the user can select the SCV fitting their requirements as well as the number of iterations, and whether a community detection algorithm is to be applied or not.

The reduced graphs of all iterations are visualized in a "Graph Grid" tab such that the user is made aware of the different stages of complexity reduction and the respective implications.

VI. Evaluation

In a simulation study, we evaluated the performance and effectiveness of our proposed technique to reduce the visual complexity of data sets. In the next section, we provide insights into our experimental setup and details of our simulation study. After that, we present and discuss our results for different network classes.

A. Experimental Setup

In this section, we describe and discuss our selection of data sets, followed by the description of our simulation setup.

a) Networks: To provide useful insights, we decided to perform our simulation based on samples of real-world networks. We imagine the usage of our technique in applications based on social networks (with respect to to social ties), in (scientific) libraries (with respect to to co-authorship), biological data (with respect to protein interactions), and computer networks (with respect to to the logical interconnections) as described in Table I. We restrict our simulation to the largest connected component of the data set, as accounting for importance and relationship of non-related information is out of scope of this paper.

As the representative for a social network, we decided to use a sample of Facebook's social graph [3]. This data set (FB)

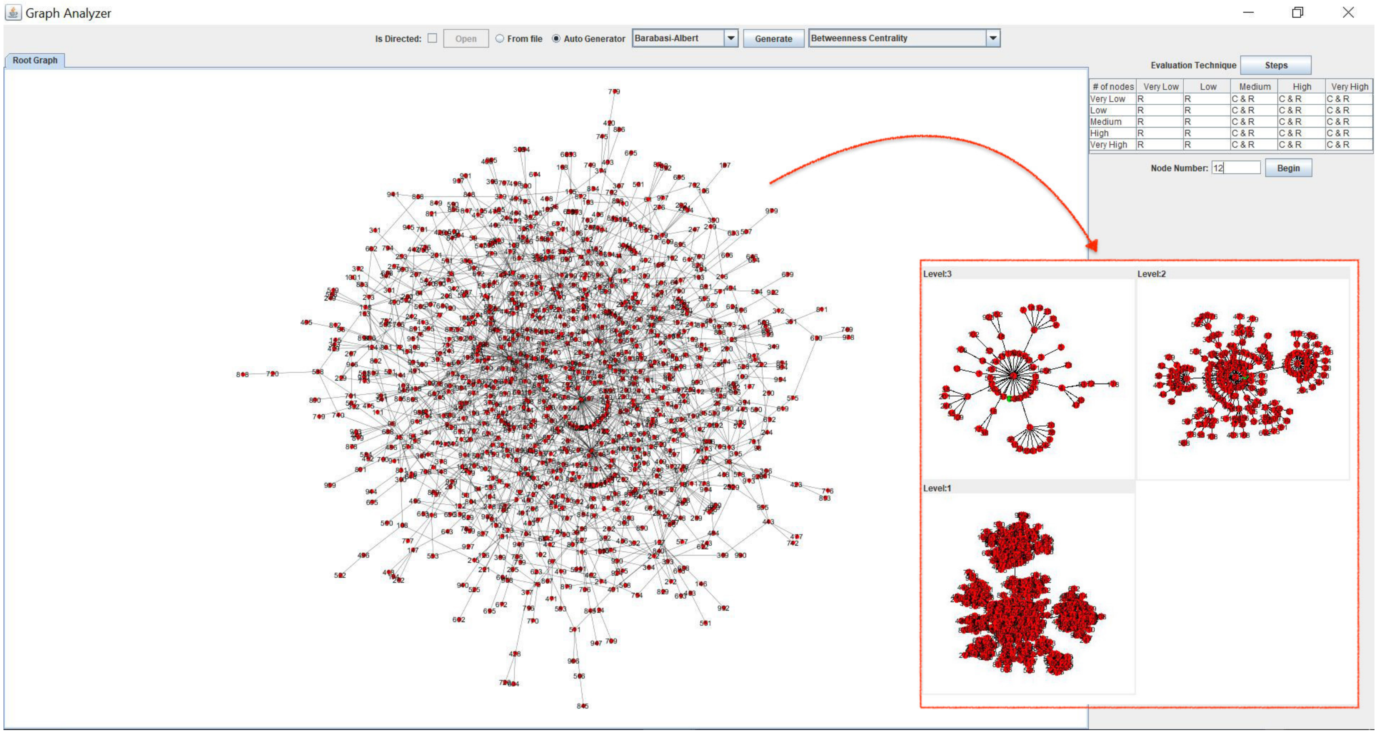


Fig. 5: Framework after loading the initial Dataset. Complexity reduction mechanism can be configured visually.

TABLE I: Large Network data sets

Class	data set	Vertices	Connections
Social	Facebook (FB)	4,029	88,234
Library	Arxiv gr-qc (CA)	4,158	13,428
Biological	Vidal (VI)	3,133	6,726
	Moreno (MO)	1,870	2,277
Computer	p2p-Gnutella08 (P8)	6,301	20,777
	p2p-Gnutella09 (P9)	8,114	26,013

consists of 4,029 vertices and 88,234 connections, representing users of Facebook and their respective interconnections.

As the representatives for library-based and collaboration-based data sets, we selected a co-authorship data set [3] that reflects the research collaborations and co-authorships in the Arxiv gr-qc (General Relativity and Quantum Cosmology) area. This data set (CA) consists of 4,158 vertices and 13,428 connections in their largest connected component.

As the representatives of biological networks, we selected the protein interaction networks vidal [31] and moreno [32]. These data sets reflect interactions of proteins on a molecular level. The vidal data set (VI) consists of 3,133 vertices and 6,726 connections; the moreno data set (MO) consists of 1,870 vertices and 2,277 connections.

As the representative of computer networks, we selected a sample of the Gnutella network [3] that reflects the interconnections of users in the Gnutella network. The sample (P8) is sampled on August 8th, 2002 (p2p-Gnutella08) and consists of 6,301 vertices and 20,777 connections. The second sample (P9) is sampled one day later on August 09th, 2002 (p2p-Gnutella09) and consists of 8,114 vertices and 26,013 connections.

b) Simulation Setup: We perform a reduction of visual complexity on each of the aforementioned data sets and calculate graph-theoretic properties that are also used as SCVs, namely closeness centrality (C_C), betweenness centrality (C_B), PageRank (PR), degree centrality (C_D), and clustering coefficient (CC).

We perform the complexity reduction on each of the data sets with each pairwise combination of selection critera values (SCV) and community detection algorithms (CDAs) been described in Sections III-B and III-C and compare and interpret the results represented by the statistical mean values and number of removed vertices.

We need a user interest expressed as one "preselected" piece of information from the data set to apply our technique to reduce visual complexity. To account for this, we selected a random vertex as user interest and repeat this whole process 15-times.

In Table II, we summarize the details of our simulation study and used abbreviations.

B. Results

In this section, we present the results of our simulative study. The results are structured according to the class of networks as aforementioned. Detailed measurements are shown in Table III.

TABLE II: Simulation Details

Property	Value(s)
SCV	Closeness Centrality (C_C) Betweenness Centrality (C_B) PageRank (PR) Degree Centrality (C_D) Clustering Coefficient (CC)
CDA	Hierarchical Clustering (HC) Direct Clustering (DC) Original Louvain (OL) Louvain w/ multilevel refinement (LM)
Repetitions	15
Iterations	3
User Interest	V.getRandomVertex()

a) Social Networks: Social networks, here the FB data set, exhibit a strongly connected core, the so-called *rich club*. These vertices are connectors between various clusters. This structural behavior leads to comparable short paths, i.e., the closeness centrality is comparably high, and core vertices exhibit a high betweenness and degree centrality. Moreover, due to the "clustering" of vertices, the clustering coefficient is also high (in the FB data set: 0.6055).

Our complexity reduction technique reduces up to 4,035 (99.9%) vertices in the three iterations. All combinations but one of the SCV and CDA combinations preserve the core of the data set, which is indicated by the high C_B, C_C, and C_D values. These indicate that the retained information is representative for their clusters. The drop of PR and CC indicate that the preserved structure roughly follows a star-topology, which is also shown in Figure 3.

Using CC as only SCV leads to a different result where different information is retained, yet, this is expected. The core vertices have by definition a lower CC value and are, thus, not likely to be above the threshold. These vertices are connected to a multitude of different other vertices that are forming their own, smaller clusters.

b) Library/Collaboration Data Set: A collaboration, or library, data set yields a similar structure to a social network. Collaborators establish clusters and an inner core of highly influential people and publications. The gr-qc data set reveals a high CC (0.5296) and a strongly connected core. While the diameter of the data set suggests otherwise, the core of the data set still shows a high connectivity and tendency towards short paths. That is evident when comparing the diameter (17) and the diameter when only 90% of vertices have to be connected (0.9-percentile effective diameter: 7.6).

Our complexity reduction technique removes up to $4,196$ vertices in the three performed iterations. We can see that most combinations of SCV and CDA perform similarly, producing star-topology-like results retaining relevant information of the core of the data set as indicated by higher C_B and C_D. The drop of the CC supports the star-topology again.

However, using CC or C_D as SCV produces different results. Using one of these properties as SCV retains information revealing the significantly higher CC, C_D, and PR and lower C_D values. Thus, retaining representatives of smaller but better and tighter connected clusters.

c) Biological Data Set: Protein interaction networks reveal a different structure, the spread between average and maximum degree is between factor 23 and 30; this factor is by a magnitude smaller than in the previously discussed social networks. This drop in degree goes hand in hand with increased diameter (13–19) and a massive decline in the CC (0.035–0.055 compared to 0.6055 on the Facebook data set).

Our complexity reduction technique removes up to 3,125 (99.74%) vertices on the vidal protein interaction network and up to 1,860 (99.47%) vertices on the moreno protein interaction network. Our technique performs similarly on both data sets, and produces star topologies regardless of the used combination of SCV and CDA. Thus, the resulting C_B is comparably high, and the C_C is very low < 0.0001. The CC also drops to 0.01–0.0004 and endorses the star topology as the remaining vertices are hardly forming local clusters. The PR is similarly low. Yet, the CO-DC combination retains local clusters, which is indicated by higher C_C and CC values.

d) Computer Networks: Computer networks, in this case the samples from the Gnutella network, are resembling either AS-networks with social-like structures on a large scale or, if consisting of only "a few" computers (compared to the whole population), random networks. As the Gnutella network were comprised of only a small subset of all Internet users, the Gnutella sample is akin to a random network. Yet, the degree distribution reveals on the base data set exponentially distributed degrees. Nonetheless, the PageRanks reveal the lack of a highly connected inner core; this core is essential to a "social" network.

Our complexity reduction techniques removed up to 6,299 (99.97%) vertices on the P8 data set and up to 8,132 (99.85%) vertices on the P9 data set. The consistently high C_B shows that some well-connected vertices are retained, but the retained vertices are distant from each other, thus, resulting in very low C_C values; only PR used as SCV preserves closer vertices. Using C_C, CC, or PR as SCV results in retaining local clusters in the results while the other SCVs do not keep clusters in their reduced data sets – this is visible by the high CC and PR values for these SCVs.

C. Destructive vs. Non-Destructive Complexity Reduction

The complexity reduction can be performed in two variations: a destructive and a non-destructive one.

First, the execution can be *destructive* by removing vertices that are not matching the user's interest, i.e., removing vertices when their SCV is lower than the threshold and if they are not selected to be a representative. During execution, the data set will shrink and, thus, free memory and reduce the computational complexity of the calculation of SCVs with each performed complexity reduction iteration. However, the user cannot revert the reduction process to earlier stages of complexity reduction.

TABLE III: Properties of Complexity-reduced Data Sets of Facebook, Collaboration, Gnutella, and the Protein Interaction Networks.

	Technique	Mean Values				
		C_B	C_C	CC	C_D	PR
FB	C_C & DC	2211000	0.3457	0.00038	298.1	0.00215
	C_C & HC	2356000	0.3478	0.00033	311.9	0.00230
	C_C & OL	1925000	0.3425	0.00037	260.3	0.00191
	C_C & LM	1925000	0.3425	0.00037	260.3	0.00191
	C_B & DC	1453000	0.3327	0.00148	205.9	0.00163
	C_B & HC	1685000	0.3182	0.00604	219.5	0.00209
	C_B & OL	2112000	0.3397	0.00291	251	0.00197
	C_B & LM	2253000	0.3389	0.00356	262.5	0.0021
	PR & DC	1520000	0.3308	0.00268	203.9	0.0017
	PR & HC	1754000	0.3356	0.00208	223.5	0.00178
	PR & OL	1588000	0.3332	0.00284	206.2	0.00163
	PR & LM	1809000	0.3339	0.00336	232.6	0.00185
	C_D & DC	1539000	0.3304	0.0024	219.7	0.00175
	C_D & HC	1810000	0.3291	0.00492	236.4	0.002
	C_D & OL	1943000	0.3311	0.0063	236.2	0.00194
	C_D & LM	2100000	0.3356	0.00451	262.3	0.00211
	CC & DC	70010	0.2638	0.07401	14.13	0.00013
	CC & HC	79330	0.2639	0.08139	15.11	0.00014
	CC & OL	30171	0.27314	2.42765	36.62	2.0282
	CC & LM	83490	0.2641	0.08111	15.71	0.00015
CA	C_C & DC	73470	0.00005	0.02978	16.36	0.00041
	C_C & HC	71590	0.00005	0.04217	8.94	0.00031
	C_C & OL	74020	0.00005	0.02445	17.66	0.00044
	C_C & LM	73470	0.00005	0.02978	16.36	0.00041
	C_B & DC	233900	0.00005	0.00592	16.95	0.00058
	C_B & HC	224500	0.00005	0.00579	16.47	0.00057
	C_B & OL	158000	0.00005	0.01006	14.12	0.00049
	C_B & LM	151600	0.00005	0.00884	13.93	0.00047
	PR & DC	222000	0.00005	0.0015	20.85	0.00067
	PR & HC	98330	0.00005	0.00811	14.92	0.00045
	PR & OL	215300	0.00005	0.00445	19.91	0.00063
	PR & LM	226500	0.00005	0.00266	20.64	0.00066
	C_D & DC	28418	4.44	1.57434	8.38	0.88412
	C_D & HC	28261	4.44	1.57331	8.37	0.88485
	C_D & OL	28370	4.44	1.57288	8.37	0.88744
	C_D & LM	97370	0.00004	0.00991	12.61	0.00039
	CC & DC	22449	4.17	0.84129	4.83	2.43779
	CC & HC	22449	4.17	0.84129	4.83	2.43779
	CC & OL	21060	4.14	0.80191	4.63	2.58936
	CC & LM	43790	0.00003	0.1903	7	0.0003
P8	C_C & DC	38492	0.00023	1.42208	9.55	3.59775
	C_C & HC	39680	0.00023	0.00003	9.53	0
	C_C & OL	38492	0.00023	1.42208	9.55	3.59775
	C_C & LM	38666	0.00023	1.43795	9.54	3.59132
	C_B & DC	82970	0.00023	0	9.92	0
	C_B & HC	88810	0.00023	0.00003	9.72	0
	C_B & OL	162900	0.00023	0	9.84	0
	C_B & LM	137000	0.00023	0	9.83	0
	PR & DC	20920	0.61077	0.58976	3.77	4.02269
	PR & HC	20920	0.61077	0.58976	3.77	4.02269
	PR & OL	18941	0.61688	0.57661	3.69	4.09170
	PR & LM	21934	0.60621	0.59507	3.81	4.14545
	C_D & DC	119200	0.00023	0	12.68	0
	C_D & HC	122900	0.00023	0	12.55	0
	C_D & OL	47420	0.00023	0.00002	10.18	0
	C_D & LM	47420	0.00023	0.00002	10.18	0
	CC & DC	46280	0.00393	2.44391	9.69	3.76677
	CC & HC	46304	0.00393	2.44341	9.69	3.76735
	CC & OL	46280	0.00393	2.44391	9.69	3.76677
	CC & LM	46576	0.00388	2.43269	9.69	3.75996

	Technique	Mean Values				
		C_B	C_C	CC	C_D	PR
VI	C_C & DC	57420	0.00009	0.00196	15.28	0.00107
	C_C & HC	57420	0.00009	0.00196	15.28	0.00107
	C_C & OL	58990	0.00009	0.00154	15.73	0.0011
	C_C & LM	54320	0.00009	0.00232	14.61	0.00102
	C_B & DC	61750	0.00009	0.00045	15.6	0.0011
	C_B & HC	58410	0.00009	0.00044	15.06	0.00106
	C_B & OL	57790	0.00009	0.00046	14.98	0.00106
	C_B & LM	56350	0.00009	0.00046	14.78	0.00104
	PR & DC	62850	0.00009	0.00600	15.86	0.00112
	PR & HC	138900	0.00009	0.00133	29.01	0.00203
	PR & OL	49980	0.00009	0.00686	13.65	0.00097
	PR & LM	85260	0.00006	0.01049	18.23	0.00136
	C_D & DC	31410	0.00009	0.02806	9.053	0.00064
	C_D & HC	156200	0.00009	0.00314	31.63	0.00221
	C_D & OL	83280	0.00007	0.04043	18.82	0.00136
	C_D & LM	111300	0.00007	0.04704	24.04	0.00172
	CC & DC	23410	8.61693	1.67212	7.98	0.00318
	CC & HC	30480	0.00008	0.1194	8.65	0.00063
	CC & OL	23590	0.00009	0.07014	7.94	0.00057
	CC & LM	24450	0.00009	0.08001	8.03	0.00058
MO	C_C & DC	33615	0.00005	0.00681	6.31	0.00115
	C_C & HC	31886	0.00005	0.00596	6.16	0.00114
	C_C & OL	33467	0.00005	0.00668	6.33	0.00116
	C_C & LM	42355	0.00005	0.0124	7.32	0.00132
	C_B & DC	22275	0.00004	0.00152	4.99	0.00096
	C_B & HC	21245	0.00004	0.00153	4.91	0.00094
	C_B & OL	62989	0.00004	0.00306	9.87	0.00178
	C_B & LM	61556	0.00004	0.00312	9.7	0.00175
	PR & DC	33375	0.00004	0.00595	6.67	0.00125
	PR & HC	30472	0.00004	0.00604	6.34	0.0012
	PR & OL	32441	0.00004	0.0062	6.51	0.00123
	PR & LM	24884	0.00004	0.00618	5.70	0.00109
	C_D & DC	25683	0.00004	0.00277	5.82	0.00109
	C_D & HC	23793	0.00004	0.0028	5.64	0.00106
	C_D & OL	24978	0.00004	0.00284	5.80	0.00108
	C_D & LM	61071	0.00004	0.00376	10.53	0.00189
	CC & DC	20061	0.00004	0.0867	5.28	0.00099
	CC & HC	20061	0.00004	0.0867	5.28	0.00099
	CC & OL	20503	0.00004	0.0857	5.33	0.001
	CC & LM	21185	0.00004	0.0845	5.39	0.00101
P9	C_C & DC	79094	0.00034	0.00133	10.14	0
	C_C & HC	79094	0.00034	0.00133	10.14	0
	C_C & OL	79094	0.00034	0.00133	10.14	0
	C_C & LM	79094	0.00034	0.00133	10.14	0
	C_B & DC	78569	0.00034	0.00001	9.73	0
	C_B & HC	115689	0.00034	0	9.93	0
	C_B & OL	249905	0.00034	0.00002	10.92	0
	C_B & LM	267735	0.00034	0.00003	11.08	0
	PR & DC	28406	0.00014	0.0001	3.69	0
	PR & HC	25826	0.00013	0.0001	3.60	0
	PR & OL	50534	0.00017	0	4.68	0
	PR & LM	46586	0.00016	0	4.60	0
	C_D & DC	145848	0.00034	0	13.17	0
	C_D & HC	145848	0.00034	0	13.17	0
	C_D & OL	163656	0.00034	0.00001	13.18	0
	C_D & LM	134312	0.00034	0	13.16	0
	CC & DC	68933	0.00034	0.00188	9.96	0
	CC & HC	68933	0.00034	0.00188	9.96	0
	CC & OL	69376	0.00034	0.00179	9.94	0
	CC & LM	71280	0.00034	0.00209	9.98	0

Second, the execution can be *non-destructive* by, e.g., concealing "removed" vertices only without deleting them from the actual graph. This variation allows moving back and forth between different stages of complexity reduction. Moreover, this also allows a user to adapt her interest, i.e., she can use gained knowledge and re-run the visual complexity reduction with a different focus/interest.

The main functionality of our proposed approach to visual complexity reduction remains unaltered by this design decision as it only affects the ability to move through the different stages of complexity reduction and the visualization for the human user.

Our proposed technique to reduce the visual complexity is capable of removing the vast majority of information in just a few iterations as we have shown before. In order to overcome unwated loss of information during the process of data reduction, our approach always allows the user to track back and select different vertices as user interest. This, however, is infeasible for dynamic data sets, which can be adressed, for example, by caching results from earlier iterations.

VII. Conclusion and Future Work

The collection and processing of large data sets got common with the rise of *big data* and powerful computing devices. Human users are hardly able to keep up with the increasing pace of collecting and accruing data. Accessing and—more important—exploring as well as understanding these data sets becomes difficult.

In this article, we present our framework that combines an interactive, user-guided approach to the exploration of large data sets with automated complexity reduction based on graph-theoretic properties and community detection. This approach reduces visual complexity in order to render visualizations of data sets more usable for human users. The reduction of visual complexity is achieved by removing parts of the data sets in a semi-automated fashion: based on a user-selected information, i.e., a selected vertex becomes a template for a high *selection criteria value* (SCV), other vertexes with a lower SCV as well as their connections are hidden from the data set. Thereby, the complexity is reduced, which we assume to render the visualization easier to comprehend. The computation of SCVs is based on graph-theoretic properties like closeness and the detection of communities within the data set.

The presented framework combines this automated computation and removal with interactive selection of a new template vertex before each complexity reduction iteration.

Our simulation study has shown that our combination of graph-theoretic properties, measuring the importance of data in the data set, and community detection, grouping similar data in the data set, is able to reduce the visual cluttering of information efficiently. If performed in a non-destructive setting, i.e., if discarded data is only concealed and not removed from the data set, human users can shift their focus when inspecting a data set to account for new insights gained by the visual inspection of data sets. The complexity reduction can then be performed again with a focus on new, shifted interests.

Our proposed technique opens an additional direction of research to support user-centric systems in rising exposure to big data and accruing amounts of data. Future research may include user studies to select the—per application field—appropriate graph-theoretic properties in order to achieve an appropriate visual complexity reduction.

Acknowledgements

This work has been co-funded by the DFG as part of project B.2 within the RTG 2050 Privacy and Trust for Mobile Users and by the German Federal Ministry for Education and Research (BMBF, Software Campus project KomBi, grant no. 01IS12054).

References

[1] T. Grube, F. Volk, M. Mühlhäuser, S. Bhairav, V. Sachidananda, and Y. Elovici, “Complexity Reduction in Graphs: A user Centric Approach to Graph Exploration,” in *International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC)*. IARIA, 2017.

[2] G. E. Moore, “Readings in computer architecture,” M. D. Hill, N. P. Jouppi, and G. S. Sohi, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, ch. Cramming More Components Onto Integrated Circuits, pp. 56–59.

[3] J. Leskovec and A. Krevl, “SNAP Datasets: Stanford large network dataset collection,” http://snap.stanford.edu/data, Jun. 2014.

[4] D. Kimelman, B. Leban, T. Roth, and D. Zernik, “Reduction of visual complexity in dynamic graphs,” in *International Symposium on Graph Drawing*. Springer, 1994, pp. 218–225.

[5] D. Holten and J. J. Van Wijk, “Force-directed edge bundling for graph visualization,” in *Computer graphics forum*, vol. 28, no. 3. Wiley Online Library, 2009, pp. 983–990.

[6] M. Sarkar and M. H. Brown, “Graphical fisheye views of graphs,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1992, pp. 83–91.

[7] E. R. Gansner, Y. Koren, and S. C. North, “Topological fisheye views for visualizing large graphs,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 4, pp. 457–468, 2005.

[8] G. W. Furnas, “Generalized fisheye views,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’86. New York, NY, USA: ACM, 1986, pp. 16–23.

[9] J. Abello, S. G. Kobourov, and R. Yusufov, “Visualizing large graphs with compound-fisheye views and treemaps,” in *International Symposium on Graph Drawing*. Springer, 2004, pp. 431–441.

[10] A. Clauset, M. E. J. Newman, and C. Moore, “Finding community structure in very large networks,” *Phys. Rev. E*, vol. 70, p. 066111, Dec 2004.

[11] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, “Network analysis in the social sciences,” *science*, vol. 323, no. 5916, pp. 892–895, 2009.

[12] M. Girvan and M. E. Newman, “Community structure in social and biological networks,” *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[13] P. Sundararajan, O. J. Mengshoel, and T. Selker, “Multi-fisheye for interactive visualization of large graphs,” in *Proceedings of the 17th AAAI Conference on Scalable Integration of Analytics and Visualization*, ser. AAAIWS’11-17. AAAI Press, 2011, pp. 46–50.

[14] V. Batagelj, A. Mrvar, and M. Zaveršnik, “Partitioning approach to visualization of large graphs,” in *International Symposium on Graph Drawing*. Springer, 1999, pp. 90–97.

[15] S. P. Borgatti, “Centrality and network flow,” *Social networks*, vol. 27, no. 1, pp. 55–71, 2005.

[16] L. C. Freeman, “Centrality in social networks conceptual clarification,” *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.

[17] E. M. Daly and M. Haahr, “Social network analysis for routing in disconnected delay-tolerant manets,” in *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2007, pp. 32–40.

[18] S. Brin and L. Page, “Reprint of: The anatomy of a large-scale hypertextual web search engine,” *Computer networks*, vol. 56, no. 18, pp. 3825–3833, 2012.

[19] “Pagerank implementation description,” http://www.ccs.northeastern.edu/home/daikeshi/notes/PageRank.pdf, accessed: 2018-05-28.

[20] R. Zafarani, M. A. Abbasi, and H. Liu, *Social media mining: an introduction*. Cambridge University Press, 2014.

[21] M. Brautbar and M. Kearns, “A clustering coefficient network formation game,” in *Intl. Symp. on Algorithmic Game Theory*. Springer, 2011, pp. 224–235.

[22] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[23] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[24] M. E. J. Newman, “Modularity and community structure in networks.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 23, pp. 8577–8582, 2006.

[25] L. Waltman and N. J. van Eck, “A smart local moving algorithm for large-scale modularity-based community detection,” *The European Physical Journal B*, vol. 86, no. 11, pp. 1–14, 2013.

[26] R. Rotta and A. Noack, “Multilevel local search algorithms for modularity clustering,” *Jrnl. of Exper. Algorithmics*, vol. 16, pp. 2–3, 2011.

[27] K. Alsabti, S. Ranka, and V. Singh, “An efficient k-means clustering algorithm,” 1997.

[28] T. Grube and S. Bhairav, “Graph exploration and complexity reduction framework,” https://github.com/timgrube/GraphExplorationComplexityReduction, accessed: 2018-05-28.

[29] “Graphstream project,” http://graphstream-project.org/, accessed: 2018-05-28.

[30] D. Jurgens and K. Stevens, “S-space,” https://github.com/fozziethebeat/S-Space, accessed: 2018-05-28.

[31] “Human protein (vidal) network dataset – KONECT,” Oct. 2016, accessed: 2018-05-28. [Online]. Available: http://konect.uni-koblenz.de/networks/maayan-vidal

[32] “Protein network dataset – KONECT,” Oct. 2016, accessed: 2018-05-28. [Online]. Available: http://konect.uni-koblenz.de/networks/moreno_propro

Reporting and Analyzing Student Behavior in 3D Virtual Worlds

Aliane Loureiro Krassmann
Graduate Program of Informatics in Education
Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre/RS – Brazil
alkrassmann@gmail.com

Liane Margarida Rockenbach Tarouco
Graduate Program of Informatics in Education
Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre/RS – Brazil
liane@penta.ufrgs.br

Felipe Becker Nunes
Graduate Program of Informatics in Education
Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre/RS – Brazil
nunesfb@gmail.com

Magda Bercht
Graduate Program of Informatics in Education
Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre/RS – Brazil
bercht@inf.ufrgs.br

Tito Armando Rossi Filho
Graduate Program of Informatics in Education
Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre/RS – Brazil
rossitito@hotmail.com

***Abstract*—Virtual Worlds are open 3D environments capable of simulating a variety of educational practices, thus reducing risks and costs with physical laboratories. However, its freedom characteristic can cause dispersion and make it difficult to navigate and focus. Aiming at improving user navigation, this paper analyzes a Heads Up Display (HUD) solution that keeps sensing the locations visited by the user, and presents it in the form of a "heat map". The main contribution has two sides: 1) a reporting system is presented, which allows teachers to easily keep up with students' behavior in the environment; 2) an analysis of experimental data has been performed, focusing on the extent and intensity of content observation, with the purpose of presenting the potential of the HUD to act as a guidance tool in educational Virtual Worlds.**

Keywords-virtual worlds; heads up display; student interaction.

I. Introduction

This paper is an extended version of [1], which proposed the use of Heads Up Display (HUD) device working as a heat map, to create a context-aware Virtual World (VW), coming to the conclusion that it helped on students' navigation. The HUD could be attached to any user's screen by touch, and it would keep dynamically inferring its context, changing colors according to the locations visited and the frequency (or duration) of visitation.

VW are 3D open environments where users, represented by their own avatars, can move around, meet and interact with other avatars [2]. This technology is capable of simulating whole environments, thus greatly decreasing costs and risks of conducting experiments in physical laboratories [3].

Also, according to Englund [4], it enables to move away from traditional classroom learning and to design activities with an emphasis not on "learning about", but with focus on "learning by being" or "learning by doing". Mastrokoukou and Fokides [5] highlight that although the development of these environments is quite a lengthy process, long term benefits of its use may arise, since it can be reused several times.

Chow [6] explains that a VW is characterized by being more exploratory, active and participatory, rather than centralized in listening and absorbing information. Multiusers are encouraged to navigate freely through different locations within the environment, to develop their own learning processes, according to their own demand [7].

Due to the great diversity of didactic materials that can be inserted in VW, such as texts, images, videos and presentation slides, different spaces or rooms can be created, separated by, for example, subject or type/category of content. Simsek and Can [8] assert that giving students the freedom to choose the type of learning material to explore, makes them more active individuals in their learning process.

As the ones who spend more time inside it tend to interact more with educational objects [9], it is desirable for the user to spend long times navigating the VW. Studies have described that the total number of visits to a learning object may indicate student interest [10].

However, these characteristics of discovery-based environments, which offer the greatest potential to promote learning, at the same time represent the biggest challenge: the complexity of the learning experience [11]. According to Mayer [12], environments designed to make users discover materials completely on their own are harder to use. Gütl [13] discovered that if users spend too much time learning to use a complex interface they might leave the environment.

This leads to several educational implications, as engagement hindrance or demotivation, as users may be distracted by numerous functions and simulated scenes in the VW, disturbing their attention from the learning contents [3].

For example, Ijaz, Bogdanovych and Trescak [14] found, in an unsupervised experiment, that participants did not properly explore the VW: important areas were left unexplored. Also, some students faced technical issues and did not fully achieve the learning objective; in consequence, scored low in the exam. In their study, Griol, Molina and Callejas [15] found that the amount of options and tools available in the VW had a negative effect in some students, who felt disoriented. They suggest this happened especially to players of video games, as they are used to follow a script, with predefined outcomes.

This way, Christensen, Maraunchak and Stefanelli [16] suggest that the educational use of VW must be carefully articulated and organized by the teacher. Csikszentmihalyi [17] emphasizes that activities stimulate more flow if they embody certain rules and clearly state what the users should do.

Moreover, for immersive learning environments to act as it is expected, with students having great flexibility while faced with numerous learning opportunities, intelligent real-time support and guidance is required [14]. This means that, in accordance with Tüzün and Özdinç [18], being informed about the objects and locations in the environment plays a key role in getting to know about other paths. Baydas et al. [19] exemplified how reflective guidance, directive signs, symbols, footprints and notice boards within the VW can be helpful in achieving the learning goals, getting some positive results.

However, one of the most complicated aspects inside VW is precisely automatic scaffolding or guidance, since it is very large and flexible [18]. As stressed by Soliman and Guetl [20], it is not unusual to see the lack of autonomous support in VW, and learners find themselves alone with no guidance.

The usual approach to provide automatic guidance on VW is concentrated on virtual agents. According to Johnson, Rickel and Lester [21] animated agents that can serve as guides are an important instructional aid, so students won't become disoriented and lost in virtual reality environments. Xie and Luo [22] identified that students performed better and in less time because they could find the destinations quickly with the help of virtual agents. Also, they felt more satisfied.

In the last decades, ubiquitous computing techniques have been used aimed at improving student performance, and consequently, beneficiating institutional cost effectiveness [23]. In line with this trend, and as an alternative from the virtual agents' scenario, several researches have benefited from the use of HUD device capabilities, available to implement in most of the VW platforms, as OpenSimulator and Second Life, to personalize and dynamize VW in different ways.

Extending the work of [1], in the current research we intend to: a) establish a way to enable teachers to easily extract reports of student's behavior in the environment, in order to help monitoring educational activities; and b) analyze differently the records from the experiment previously performed, to verify if the HUD can influence in the extent and in the intensity of didactic content observation in the environment. Thus, we will try to answer the following main research question: ***how is it possible to (1) report and (2) analyze student behavior inside Virtual Worlds?*** We will discuss on how these two aspects can be useful for student assessment and pedagogical improvements in educational VW.

The article is organized as follows: Section II presents the related work; Section III explains the research method; Section IV presents the developed report system; Section V shows the analysis performed with the data, discussing results found; ending with Section VI, which presents the conclusions and future research.

II. Related Work

In this section, we present some examples of how HUD and heat map devices have been used towards optimizing user experience.

Using Second Life platform, Shah, Bell and Sukthankar [24] implemented a recommendation system that suggests places to visit, personalized with the user's destination preferences. To acquire data on users' travel patterns, they developed a custom tracker object using the Linden Scripting Language (LSL), which periodically prompts the user to enter information describing and evaluating its current location. The HUD could be worn on the right or left side of the screen, and it monitor the avatar current location (x, y, z coordinates).

In the study of Keelan et al. [25] each participant was given a HUD that allowed them to indicate up to eight emotional states throughout the Second Life visitation, four positive and four negative. The goal was to identify and analyze which aspects invoked emotional responses, and what kinds of information were considered trustworthy or untrustworthy by the users.

One of the promising VW applications that seem to demand such a device is the visit of virtual museums. In this field, Sookhanaphibarn and Thawonmas [26] emphasize that personalization can play a key role for increasing the number of return visitors. They mention the idea of using HUD to show personalized recommendations in VW, like to what has already been implemented for physical museums, but with the advantages of requesting simple implementation and no additional cost.

Likewise, Ward and Sonneborn [27] implemented a HUD that provides subtitles of dialogue in many languages, including English, French, German, Spanish, Italian, and

Portuguese. Also, the HUD records the avatar's position so users can know where they are in the build and receive audio, pictorial and textual information about what they are seeing on their visit.

In the field of displaying or highlighting locations to situate the student, the Virtual Learning Environment (VLE) MOODLE in more recent versions (2.7 onwards) has a plugin that implements the "heat map" concept to help user navigation, using a color scheme (yellow, orange and red) to represent the "heating". The heat map highlights areas as well as components that highly attracted students' attention by counting the number of clicks.

Rakoczi [28] developed a similar system, but tracking the student's eye movements to identify areas "more looked", that is, which captured more attention, as shown in Figure 1.

Figure 1. Heat map of MOODLE's calendar [28].

Similarly to the studies presented, Krassmann et al. [1] showed the initial results of the development and application of a context-aware HUD that works as a heat map, "heating" (changing colors) as the locations in the VW are visited more often, allowing the user to be aware of his/her own navigation behavior. The authors explored how this device could help on improving engagement and interaction time, getting some good results.

In this extended version, we present the development of a system that allows teachers to extract reports from user's behavior in the Virtual World, enabling to see patterns of students' access in a simple and intuitive way, facilitating monitoring and assessment, without the need for the teacher to be in person observing the activity. Also, we analyze some of the data gathered from the experiment performed in [1], using statistical techniques to investigate new hypothesis, as the impact of the HUD in the extent and intensity of content observation, and in face of the difficulty level of accessing locations in the virtual environment.

III. MATERIALS AND METHOD

The research is an exploratory quasi-experimental study, with a convenient sample of individuals who had a minimum level of computer skills. A region in the Virtual World from AVATAR Project [29] (an acronym in Portuguese that means Virtual Learning Environment and Remote Academic Work) was used. This project intends to investigate, test and promote the training and the use of virtual laboratories in immersive environments, using the open-source OpenSimulator platform. A set of regions are implemented in this VW, within different subjects as Physics, Electricity, Chemistry, among others.

Students enrolled in courses at the authors' university and colleagues that work on the project were invited to participate in the experiment, which took place on the university facilities. Singularity [30] and Firestorm [31] viewers were used to enter the VW, being those free softwares capable of rendering the 3D graphical environment appropriately.

To explain more in detail the method of research, we have created subsections regarding the VW development, the heat map HUD implementation and the experiment design and execution.

A. The Virtual World

Two virtual laboratories, in one region from AVATAR Project, were used: Waves and Wireless Networks, which are introduced on Krassmann et al. [32]. According to the authors, in these environments students from Secondary or Technical education have the opportunity to visualize the practical side of some abstract concepts that are part of their daily life, as the propagation of radio waves and Wi-Fi communication.

To improve the organization of content and provide a clear distribution, these two teaching contents were divide into 12 topics, distributed along 12 specific locations in the two laboratories, which are presented in Table I.

TABLE I. LOCATIONS AND TOPICS DISTRIBUTION

Number	Content topic/subject
1	Wave Characteristics
2	AM and FM Radio Waves
3	Wave Phenomena
4	Electromagnetic Spectrum
5	Introduction to Wireless Networks I
6	Introduction to Wireless Networks II
7	Wireless Networks Topologies
8	Infrared and Bluetooth
9	Range of Wireless Networks I
10	Range of Wireless Networks II
11	Range of Wireless Networks III
12	Material interference in propagating wireless networks

In each of the 12 locations, different didactic materials such as videos, presentation slides, images, texts, animated digital media, audios, web pages embedded in QR (Quick Response) code and simulations were implemented, according to the location subject. Each one was identified as to their type through luminous plaques and with short instructions displayed on the bottom, seeking to provide an intuitive and independent navigation for users.

Figure 2 presents a screenshot of the environment entrance. It can be seen that a control panel was placed at the beginning of the path, containing buttons that give the user introductory information, for example, on how to move around, use the didactic materials and what is expected in this visitation. Also, there's a green button where the user can attach itself the HUD. In this case, the user have already pushed the button and the HUD is on the right side the screen.

Figure 2. Waves and Wireless Networks Laboratories entrance.

To help on student guidance, Figure 3 shows plaques simulating wooden material with arrows pointing the directions to each location. The aim of this signalization was to ensure that users without the HUD could also navigate the environment and visit all the locations available.

Figure 3. Plaques with arrows pointing the directions to each location.

B. The heat map HUD

The HUD is composed of a numerical map of the environment. It can be attached by the user by touching one time on the corresponding button, and it goes to its top right screen. It works sensing the places visited, recording into a database and retrieving this information in real time.

The HUD is composed of 12 prims (primitives – 3D object unit) that are linked. Each prim has its own texture that identifies the number and the topic of one location, and its own scripts to change its color from white to red, "heating" according to the frequency of access or time elapsed on visitation. To do this, the classification presented in Table II was idealized and adopted.

TABLE II. HEAT MAP CLASSIFICATION ON HUD

Color	Tag	Frequency	Time elapsed
Yellow	Weak	Second time	2 minutes
Orange	Medium	Third time	3 minutes
Red	Strong	Fourth time	4 minutes

In this sense, if the user has remained at one of the 12 locations for at least two minutes, according to our classification, it is assumed he/she is visiting this location for the second time, so the prim on the HUD that represents this location turns yellow; after three minutes it turns orange, and after four minutes it turns red. The sensor checks every five seconds for the presence of an avatar and consults in the database if this specific user has already visited the place, registering it only after every minute (60 seconds), in an incremental way. The prim also has the ability, by touching it, to tele transport users to the respective location.

Figure 4 shows the heat map following the user's avatar at two different locations in the VW. It can be seen, for example, that this user has visited Location 5 very often (red), but Location 3 lacks visiting, as it is still white (top screen). It also shows that Location 8 is now changed from yellow to red (bottom screen). So it can be concluded that this user got back to that location twice more and/or has spent more two minutes in the same spot (from yellow – 2 minutes; to red – 4 minutes).

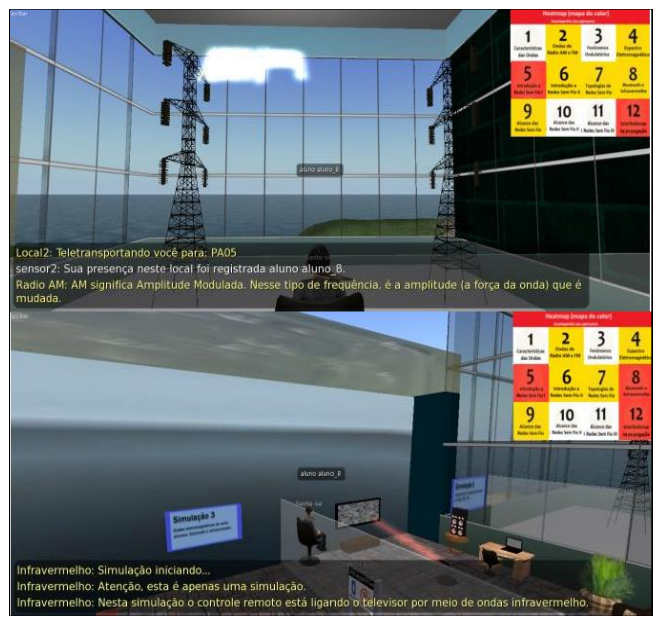


Figure 4. HUD following the student in two different locations.

The identification of locations inside the VW were signalized with circles on the floor, with sensors programmed to identify avatars presence in a radius of eight meters. Figure 5 shows a screenshot of Location 4 "Electromagnetic Spectrum", with the gray circle on the floor identifying it ("L4").

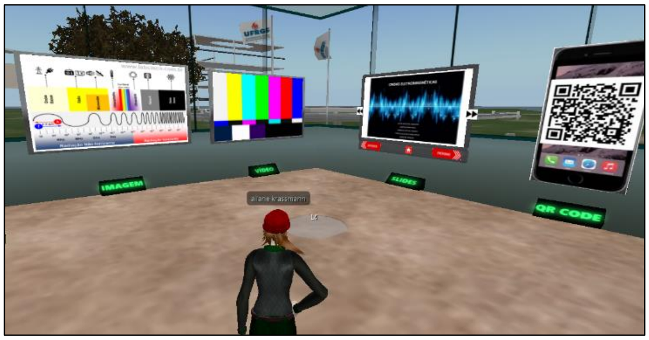


Figure 5. Screenshot of Location 4 “Electromagnetic Spectrum”.

Besides LSL programming language, OpenSimulator Scripting Language (OSSL) was used to program the HUD. To capture data from each user in real time, the sensors collect data and send it through HTTP requests. All this procedures are managed by a WAMP system (Windows, Apache, MySQL, PHP).

More specifically, the function *llSensor* was used to create an invisible monitoring field, defining the area of the sweep (eight meters). This sensor is activated every 60 seconds, through the use of *llSetTimerEvent(60)* function, whose call is carried out by the instruction. In this way, the sensor checks if there are users in its coverage range every 60 seconds. If so, the avatar name is verified and stored, along with the information about the current location.

The script which performs this sensing operation in the VW environment works in the following way (written in pseudocode).

```
1. Set the time: llSetTimerEvent(60)

2. When time reaches, activate sensor:
   llSensor("", NULL_KEY, AGENT, 8, PI)

3. When sensor activates, search for avatar name:
   avatar_name = llDetectedName(x) + avatar_name

4. Once avatar name is captured, gather the data:
   string hud =
   avatar_name+location_id+location_name+time

5. Pass the data to PHP file:
   llHTTPRequest("http://server/heatmap.php?dados
   _user="+hud, [HTTP_METHOD,"GET"],"")

6. Get the answer from server:
   http_response(key request_id, integer status,
   list metadata, string hud)
```

As shown in the pseudocode presented, the information is sent through the function *llHTTPRequest*. The function *HTTP_METHOD* “GET” is the one which enables the PHP file in the server to receive the data. The file receives it and “explodes” it in an array, separating location id, location name and time the user remained in that location. The names of each user that were inside the sensor area at the moment of data collection are separated and organized. Another function works registering time records in which this occurred. After all these treatments, data are ready to be sent to MySQL database, where they are stored.

Data are stored in the “heat map” table of the database, as follows: a) user’s name; b) name and ID of the location visited; c) time user remained in that location (60, 120, 180 or 240 seconds); d) the current heat map status for the user at that location (white, yellow, orange or red); e) time records. When visiting a location for the first time, a new complete record is inserted in this table, so time and current heat map status attributes are subsequently updated, according to an incremental analysis of time elapsed.

Immediately after this procedure, the *avatar_name* attribute has to be cleaned, so it can be received again and compared in the table. The *http_response* then brings the answer, which is passed through the channel 225 (selected by us) to update the HUD, if necessary.

This process can occur simultaneously for several users, receiving all interactions occurred from the sensors inserted inside the environment, allowing the sensing of multiple users in different locations, without competition problems.

C. *The experiment*

Data from an experiment performed in Krassmann et al. [1] is analyzed. As described by the authors, on the occasion 16 individuals, divided into control and experimental groups, were informed about the experiment goals, the voluntary aspect of their participation and the complete confidentiality of any data gathered about them. Each one received an individual login to access the 3D environment. They were instructed to freely and intuitively navigate in the VW, without any pedagogical path or visitation time previously defined. The purpose of this orientation was to provide them with freedom to interact in places they considered appropriate, visiting the desired materials and remaining in each location as long as they were interested.

Immediately after the session, that lasted for an average of approximately 33 minutes, a questionnaire was administered, containing demographic questions and items regarding the navigation in the environment, including the impressions about the heat map (HUD) for the experimental group. The main findings are presented in the original paper [1]. In this extended version we focus on creating a reporting system and analyzing student interaction in a more profound way, which corresponds to the two next sections.

IV. Reporting Student Interaction

As mentioned by Balderas et al. [33], VWs developed using open-source software — such as OpenSimulator — allow developers and teachers to access student logs and retrieve valuable information on learners’ in-world behavior and interaction. This possibility, explored in this research, is important not only in terms of assessment: it also facilitates the identification of learner profiles and VW behavior patterns, which can help improving the environment, as it allows the accomplishment of different types of actions to assist the students during their interaction [27].

Duncan, Miller and Jiang [34] corroborates that having a good understanding of the user's in-world behavior can help on improving the educational use. They suggest that, as students virtually participate, it may be hard for the teachers to monitor the educational process, being difficult to tell whether the students are actually learning or playing in the VLE.

However, a high degree of dependence on the teacher or his/her constant intervention in the activity is also undesirable. Ijaz, Bogdanovych and Trescak [14] discovered that it decreases interactivity and immersion of participants: users became less engaged in the experience and were keen to finish it quickly.

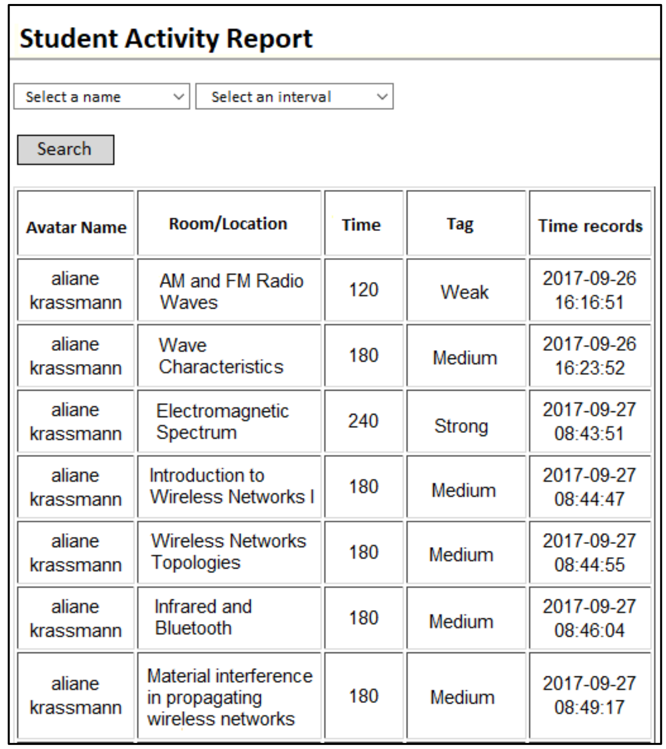
In order to improve and facilitate student monitoring and evaluation, by the pedagogic point of view and in an autonomous way, a reporting web-system was developed. Its objective is to facilitate the process of analysis of student's interactions inside the 3D VLE without the need of the teacher assistance. The system was basically developed with the PHP programming language and the connection to the MySQL database, with the previously described "heat map" table, which stores all user interactions received from the HUD.

With the data of each user, it is possible to identify which places were visited in the environment, since in these places sensors are triggered by the presence of a user and record this information in the database. In this sense, based on the data collected from students interactions in the two laboratories (12 locations), we have implemented six different types of reports, described as follows.

1. **Student Activity Report:** the data of a specific user, selected by the teacher, are presented for an individual assessment, showing locations the student visited, how long he/she remained on each one, and the tags related to each time. An example is presented in Figure 6.

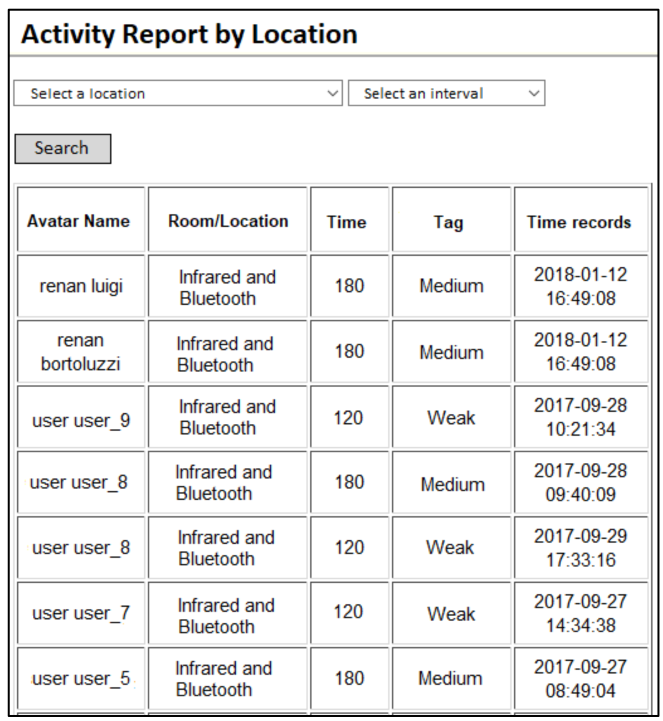
2. **Activity Report by Location:** in this report all interactions occurring in a specific location are extracted, allowing the teacher to see an overview of places that have been visited the most, and, on the other hand, which need to be more stimulated to visitation. For instance, it might be a warning for the teacher that a specific location lacks attractiveness. Figure 7 shows a clipping of an example of this report, showing all user presence captured in Location 8 "Infrared and Bluetooth".

3. **Activity Report by Time:** the teacher has the possibility to investigate, in this report, the interaction that occurred in different intervals of time, covering all locations and students. For example, it can show all the places and users with a 180 seconds tag (visited three times or for three minutes). It can be useful for the teacher, for instance, to immediately verify places more accessed (time = 240 seconds) and, on the other hand, places that were never visited (time = 0 seconds).



Avatar Name	Room/Location	Time	Tag	Time records
aliane krassmann	AM and FM Radio Waves	120	Weak	2017-09-26 16:16:51
aliane krassmann	Wave Characteristics	180	Medium	2017-09-26 16:23:52
aliane krassmann	Electromagnetic Spectrum	240	Strong	2017-09-27 08:43:51
aliane krassmann	Introduction to Wireless Networks I	180	Medium	2017-09-27 08:44:47
aliane krassmann	Wireless Networks Topologies	180	Medium	2017-09-27 08:44:55
aliane krassmann	Infrared and Bluetooth	180	Medium	2017-09-27 08:46:04
aliane krassmann	Material interference in propagating wireless networks	180	Medium	2017-09-27 08:49:17

Figure 6. Example of User Activity Report (Report 1).



Avatar Name	Room/Location	Time	Tag	Time records
renan luigi	Infrared and Bluetooth	180	Medium	2018-01-12 16:49:08
renan bortoluzzi	Infrared and Bluetooth	180	Medium	2018-01-12 16:49:08
user user_9	Infrared and Bluetooth	120	Weak	2017-09-28 10:21:34
user user_8	Infrared and Bluetooth	180	Medium	2017-09-28 09:40:09
user user_8	Infrared and Bluetooth	120	Weak	2017-09-29 17:33:16
user user_7	Infrared and Bluetooth	120	Weak	2017-09-27 14:34:38
user user_5	Infrared and Bluetooth	180	Medium	2017-09-27 08:49:04

Figure 7. Example of Activity Report by Location (Report 2).

4. **Activity Report by Color:** similar to the previous report, in this case it is possible to perform the analysis of student interaction according to the color tags defined in the heat map (white, yellow, orange and red), covering all

locations. In this case, teacher can use the color tag to identify places more and less visited.

5. **Detailed Activity Report:** seeking to provide the teacher with more details of each student behavior, this report presents the summarized data of a selected user, as it can be seen in the clipping of example presented in Figure 8. It shows the total number of visits in each location, how long it was the total visit duration on each place, and the number of heat map tags received for each location. In this sense, this report allows the teacher to see where a specific student have been in the VLE and for how long, allowing, for example, to evaluate him/her.

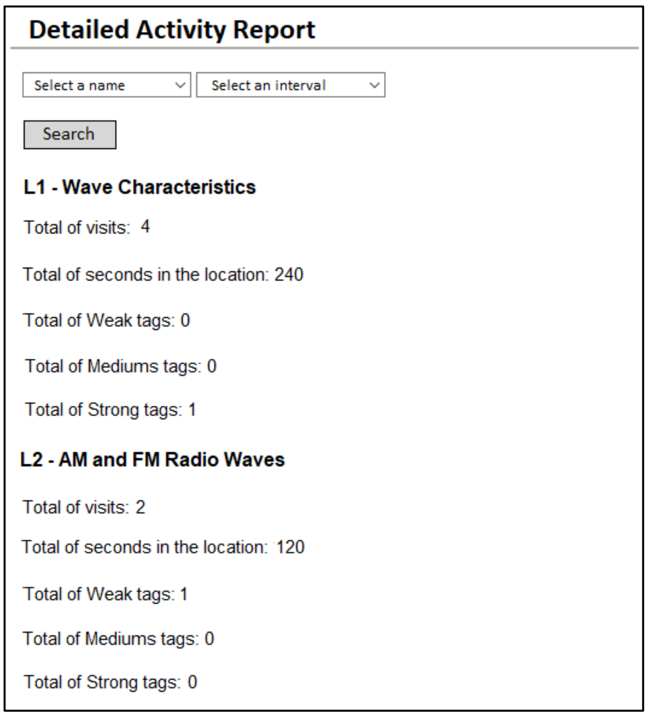


Figure 8. Example of Detailed Activity Report (Report 5).

6. **Full Report:** this report aims to provide the teacher with all the information stored in the database in a summarized way. It enables to have a panorama of the whole class interaction in the VW, and to do filtered searches in a diversity of ways, seeking to analyze more specific data. It also allows, for instance, the teacher to evaluate the activity, to analyze if all the locations projected for learning were used (or useful).

Each report is ordered decreasingly chronologically, and it can be generated in three different periods: weekly, monthly and complete, by selecting an option in field "select an interval". In addition, all the reports have the option of being exported in Portable Document Format (PDF), or can be downloaded in the form of customizable worksheets, so the teachers have the freedom to manipulate and store the reports according to their preference or need.

Through this web system, teachers can perform activities with a group of students in the Virtual World without the need of being present to observe the procedure all the time, consequently influencing in the student freedom and feelings of immersion [14]. Also, this way, giving proper instructions, students can access the environment from home.

The data are easily available on the internet, through any device. It can be used to assist the teacher on the self-evaluation, the analysis and/or the assessment of the whole class or each student individually, at any time and from anywhere.

V. Analyzing Student Interaction

This section aims to complement the analysis presented in Krassmann et al. [1], investigating data from the experiment performed by the authors in a more thorough way.

The original article did some inferences based mostly on students' self-report, that is, based on subjective opinions about the VW, which implicates on individuals personal aspects, as personality or mood. Focusing on making a more objective analysis, aiming to identify patterns from the experimental group (using the HUD), and compare it with the control group (that did not use the HUD), data from students interactions in the Virtual World were analyzed by the light of statistical inference tests, using the software Minitab version 17.

A sample composed of 10 individuals, being six from the experimental group and four from the control group was selected. This low sample is because not all participants' records were properly registered in the database, due to adjustments done in the scripts by the researchers during the experiments period.

The following new research questions (RQ) have driven this data analysis:

- RQ1 - Is there any impact of the HUD in the extension of content observation inside the Virtual World?
- RQ2 - Is there any impact of the HUD in the intensity of content observation inside the Virtual World?
- RQ3 - Is there any impact of the HUD in the difficulty level to find locations inside the Virtual World, regarding the extension and the intensity of content observation?

To investigate RQ1, it was created a formula that identifies the degree of student experimentation of all locations, which we named ECO (Extension of Content Observation), since each location relates to a topic content, calculated using (1).

$$\frac{\sum_{i=1}^{n} QuantityofAccessedLocationsbystudenti}{QuantityofAvailableLocations \times n} \quad (1)$$

n=total of assessed students

As result, the ECO of control group had an average of 0.75, while the value obtained for the experimental group was 0.66. It means that, overall, students from control group accessed 75% of locations at least once, and the students from experimental group accessed a bit less, about 65%.

Thus, we can answers RQ1 with a yes, there is some impact of the HUD in the extension of content observation in the VW, but not as we expected: the users with the heat map device visited a lower number of locations. In other words, higher level of content observation (locations) was covered by control group. Some inferences trying to explain this fact are made on RQ2 explanation. However, t-tests performed to compare ECO of both groups could not reject the null hypothesis of equality, with p=0.57 at confidence level of 95%.

RQ2 is concerned with the intensity of access of locations, considering to the heat map classification (Table II). It is here called ICO (Intensity of Content Observation), and is calculated according to (2).

$$\frac{\sum_{i=1}^{n} Quantityof4thtimeAccess(\quad)LocationVisit_i}{QuantityofAvailableLocations \times n} \quad (2)$$

n= total of assessed students

For this measure, instead of considering all locations visited by users, it was considered only locations that were accessed at least four times (or 4 minutes – strong/red tag).

As result, the ICO for the control group was 0.25 and for the experimental group 0.28. This answers RQ2, showing that yes, there is an impact, although small, of the HUD in the intensity of content observation in the VW: individuals from experimental group visited locations with a bit higher frequency. However, the null hypothesis that the ICO of both groups is equal has not been rejected, with p=0.82.

The results for RQ2 indicate a positive impact of the heat map in the Virtual World activity and complement RQ1. It shows that, although students using the heat map visited less locations, they visited it with more intensity (with more frequency or stayed for more time). Among the reasons for this fact, we speculate that maybe they felt more compelled to achieve the heat map completeness (turn red locations on the HUD).

One of the potential advantages of using this concept of heat map in VW is to highlight areas that have not been explored yet by the users, especially places that may be more difficult to find in the 3D virtual space, in consequence to the environment design. We investigated this usefulness in RQ3.

First, we assigned a difficulty level of access for each location in the Virtual World, considering researchers' experience while observing student difficulty to find them when navigating the 3D VLE. So, it was created the classification displayed in Table III.

TABLE III. DIFFICULTY CLASSIFICATION OF LOCATIONS

Location number	Difficulty to find
1, 4, 5, 6, 7, 8, 10	Easy
2, 3, 9	Moderate
11, 12	Hard

Easy locations are the ones considered more visible, with no obstacles to reach it. Moderate locations are a little hidden, requiring students to go inside rooms and/or to climb stairs. Hard locations are the ones in different buildings or in more hidden places, as behind walls, for example. Figure 9 shows an example of a moderate to find location, as the user had to climb the stair on the right side of the screen to be able to see the wireless propagation simulation (Location 9).

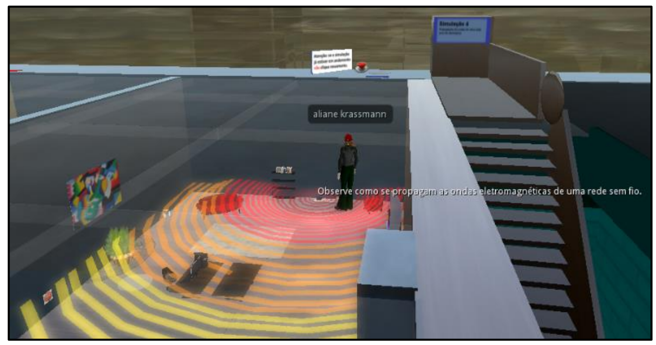


Figure 9. Screenshot of Location 9 "Range of Wireless Networks I".

Following this classification, ECO and ICO metrics were calculated for each level of difficulty and compared between control and experimental groups. The results are shown in Table IV.

TABLE IV. ECO AND ICO ACCORDING TO ACCESS DIFFICULTY

Location classification	ECO		ICO	
	Control	Exp.	Control	Exp.
Easy	0.78	0.76	0.46	0.40
Moderate	0.83	0.61	0.08	0.22
Hard	0.50	0.41	0.00	0.17

This result indicates a clear impact of the heat map, especially in ICO metric for the "moderate" and "hard" location classifications, since the difference of values between groups are bigger (0.14 and 0.17). The data show that students from experimental group accessed for much more time (with more intensity) the locations classified as hard or moderate to find than students from the control group.

Therefore, RQ3 can also be answered with a yes, there is an impact, small but positive, of the HUD in the difficulty level of access of locations in the Virtual World, regarding the intensity (frequency) of access. In this sense, again it may be inferred that the heat map have motivated students to turn locations red on the HUD, even the ones harder to find, frequenting it more. This highlights the benefit of a little of gamification implemented in the activity, as the students sought to complete their HUD while using it. Still, the t-test did not reject the null hypothesis of equality between groups.

To sum up, these new research questions (RQ1, RQ2, RQ3) showed more objective results, which could not be obtained only with students self-reports. This justifies the need and the importance of automatically keeping track of students' interactions in the VW, recording it properly into databases and making it accessible to teachers in an organized way.

However, in spite of some noticeable difference between groups, in all cases all *p*-values were higher than 0.05, not rejecting the null hypothesis at confidence level of 95%. So a caution must be taken to expand the conclusions of this research to a population of students, due to the small sample size analyzed in this study.

VI. Conclusions And Future Work

Virtual Worlds are tools that simulate the real world in 3D, providing teachers with many educational possibilities, as the creation of virtual laboratories [7]. But the lack of autonomous support, while learners are interacting in the environment, might prejudice the experience [19].

In order to contribute with this theme, we have presented a solution that takes advantage of users' context information to present a real-time heat map, using the HUD device available in most VW platforms. The goal is to help on user navigation in the VLE, highlighting places visited according to the frequency of access, keeping students aware of their activities (places visited), and at the same time, personalizing their experience. First results of this application demonstrated that it can increase engagement and interaction time [1].

In this paper we have tried to answer our main research question: ***how is it possible to (1) report and (2) analyze student behavior inside Virtual Worlds?***

To answer (1) we have introduced a web system that allows the teacher to easily extract reports from students' behavior inside the VW, retrieved from sensors implemented using LSL, OSSL and PHP programming languages. This way, the educator does not need to stay monitoring students to evaluate the activity progress, and in consequence, it minimizes possible negative influences of his/her presence on students' immersion and interactivity. Six different types of reports were constructed, allowing to see, for example, locations less visited, which may, in result, indicate the need for adjustments in the environment to improve visitation in those places.

To investigate (2), we have performed a deeper analysis of data from a previous experiment [1], to identify the usefulness of the heat map HUD in providing guidance inside the VW, in order to help the students to reach the learning objectives. Three new research questions were investigated, which using statistical techniques, demonstrated the existence of an impact of the HUD in the extension (number of places visited) and in the intensity (frequency of location visitation) of educational content observation inside the environment. Overall, the results have shown how the contribution of the gamification aspect of the HUD device motivated students to achieve the "red" tag on the heat map, even in places more difficult to find. In this sense, we have also presented some possibilities of data analysis in educational Virtual Worlds.

As the main contribution, we showed that, concerning student guidance, intelligent agents are not the only solution; other types of autonomous support may be helpful. The HUD has as advantages being flexible, as it can be attached to any particular user, and it functions individually, different from common agents that usually are not personalized for each user. In addition, the device is different from common heat maps, which highlight items or places visited by all users. All the tools used are open source, allowing researchers and educators from around the world to reproduce the same or better solutions.

As future works, regarding (1), we plan to integrate the ECO and ICO formulas to the report system, to automatically obtain the levels of student interaction in the VLE, regarding extension (the number of locations/topics visited) and intensity (the frequency of location/topic visit). In the same manner, we intend to include a function to attribute notes for student performance, measured by visit duration or coverage, allowing them to see their own reports after the experience.

Concerning (2), we will apply the VW with the HUD and the reporting system with more teachers, from different areas and environment designs, to compare the results and optimize the tool, gathering data that can be analyzed to increase the validity of results. Based on these new results, we will construct a model to improve student's unsupervised learning process inside Virtual Worlds.

References

[1] A. L. Krassmann, F. B. Nunes, T. A. Rossi Filho, L. M. R. Tarouco, and M. Bercht, "Heads Up Displays (HUD) as a Tool to Contextualize the User in 3D Virtual Worlds", The Eleventh International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2017), IARIA, Nov. 2017, pp. 169-175, ISBN: 978-1-61208-598-2.

[2] D. J. H. Burden, "Deploying embodied AI into virtual worlds", Knowledge-Based Systems vol. 22, no. 7, pp. 540-544, 2009.

[3] D. Liu, K. K. Bhagat, Y. Gao, T. W. Chang, and R. Huang, "The Potentials and Trends of Virtual Reality in Education", In Virtual, Augmented, and Mixed Realities in Education, pp. 105-130, Springer, Singapore, 2017.

[4] C. Englund, "Exploring approaches to teaching in three-dimensional virtual worlds", The International Journal of Information and Learning Technology, vol. 34(2), pp. 140-151, 2017.

[5] A. Mastrokoukou and E. Fokides, "Development and Evaluation of a 3D Virtual Environment for Teaching Solar System's Concepts", International Journal of Education and Information Technology, vol. 1(5), pp. 148-154, 2016, ISSN: 2381-7410 (Print); ISSN: 2381-7429 (Online).

[6] M. Chow, "Determinants of presence in 3D virtual worlds: A structural equation modelling analysis", Australasian Journal of Educational Technology, vol 32(1), pp. 1-18, 2016.

[7] M. Rico, J. Rodriguez, D. Riofrío-Luzcando, and M. Berrocal-lobo, "A Cost-Effective Approach for Procedural Training in Virtual Worlds", Journal of Universal Computer Science, vol. 23(2), pp. 208-232, 2017, doi: 10.3217/jucs-023-02-0208.

[8] I. Simsek,and T. Can, "The Design and Use of Educational Games in 3D Virtual Worlds", Society for Information Technology & Teacher Education International Conference, AACE, Mar. 2016, pp. 611-617.

[9] J. Cruz-Benito, R Therón, F. J. García-Peñalvo, and E. P. Lucas, "Discovering usage behaviors and engagement in an Educational Virtual World", Computers in Human Behavior, vol. 47, pp. 18-25, 2015.

[10] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an 'early warning system' for educators: A proof of concept", Computers & Education, vol. 54(2), pp. 588-599, Feb. 2010, doi: https://doi.org/10.1016/j.compedu.2009.09.008.

[11] R. Moreno, R. E. Mayer, H. A. Spires, and J. C. Lester, "The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents?", Cognition and Instruction, vol. 19(2), pp. 177-213, 2001, doi: https://doi.org/10.1207/S1532690XCI1902_02.

[12] R. E. Mayer, "Multimedia learning", Psychology of Learning and Motivation, vol. 41, pp. 85-139, 2002.

[13] C. Gütl, "The support of virtual 3D worlds for enhancing collaboration in learning settings", In Techniques for fostering collaboration in online learning communities: Theoretical and practical perspectives, pp. 278-299, 2011, doi: 10.4018/978-1-61692-898-8.ch016.

[14] K. Ijaz, A. Bogdanovych, and T. Trescak, "Virtual worlds vs books and videos in history education" Interactive Learning Environments, vol. 25(7), 904-929, 2017, doi: https://doi.org/10.1080/10494820.2016.1225099.

[15] D. Griol, J. M. Molina, and Z. Callejas, "An approach to develop intelligent learning environments by means of immersive virtual worlds", Journal of Ambient Intelligence and Smart Environments, vol. 6(2), pp. 237-255, 2014.

[16] I. Christensen, A. Maraunchak, and C. Stefanelli, "Added value of teaching in a virtual world", In The Immersive Internet, Hampshire: Palgrave Macmillan, pp. 125-137, 2013.

[17] M. Csikszentmihalyi, "Finding flow", vol. 131, New York: Basic Books, 1997.

[18] H. Tüzün and F. Özdinç. "The effects of 3D multi-user virtual environments on freshmen university students' conceptual and spatial learning and presence in departmental orientation", Computers & Education, vol. 94, pp. 228-240, 2016.

[19] O. Baydas, T. Karakus, F. B. Topu, R. Yilmaz, M. E. Ozturk, and Y. Goktas, "Retention and flow under guided and unguided learning experience in 3D virtual worlds", Computers in Human Behavior, vol. 44, pp. 96-102, 2015.

[20] M. Soliman and C. Guetl. "Intelligent pedagogical agents in immersive virtual learning environments: A review", The 33rd international convention MIPRO, IEEE, May 2010, pp. 827-832.

[21] W. L. Johnson, J. W. Rickel, and J. C. Lester, "Animated pedagogical agents: Face-to-face interaction in interactive learning environments". International Journal of Artificial Intelligence in Education, vol. 11(1), pp. 47-78, 2000.

[22] T. Xie and L. Luo, "Impact of Prompting Agents on Task Completion in the Virtual World", International Journal of Online Engineering (iJOE), vol. 13(6), pp. 35-48, 2017.

[23] V. Kellen, A. Recktenwald, and C. Bumgardner, "P–An Open Source Personalization Platform for Higher Education", University of Kentucky, Lexington-USA. [Online]. Available from: https://pdfs.semanticscholar.org/1e5b/8023cbfa9d281a2ce4ed71577d2eadc0b4ce.pdf 2018.06.09.

[24] F. Shah, P. Bell, and G. Sukthankar, "A Destination Recommendation System for Virtual Worlds", Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010), 2010, pp. 475-476.

[25] J. Keelan, L. B. Ashley, D. Morra, V. Busch, K. Atkinson, and K. Wilson, "Using virtual worlds to conduct health-related research: Lessons from two pilot stuedis in Second Life", Health Policy and Technology, vol. 4 (3), pp. 232-240, 2015, doi: https://doi.org/10.1016/j.hlpt.2015.04.004.

[26] K. Sookhanaphibarn and R. Thawonmas, "Digital Museums in 3D Virtual Environment", Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena, IGI Global, vol. 1, 2011, doi: DOI: 10.4018/978-1-60960-040-2.ch042.

[27] T. B. Ward and M. S. Sonneborn, "Creative expression in virtual worlds: Imitation, imagination, and individualized collaboration", Psychology of Popular Media Culture, vol. 1, pp. 32-47, 2011.

[28] G. Rakoczi, "Cast your eyes on moodle: An eye tracking study investigating learning with Moodle", The 4th International Conference Moodle.si, May 2010.

[29] AVATAR Project. Universidade Federal do Rio Grande do Sul. [Online]. Available from: http://www.ufrgs.br/avatar/. 2018.06.09.

[30] Singularity Viewer Official website. [Online]. Available from: http://www.singularityviewer.org/ 2018.06.09.

[31] Firestorm Viewer Official website. [Online]. Available from: http://www.firestormviewer.org/ 2018.06.09.

[32] A. L. Krassmann, T. A. Rossi Filho, L. M. R. Tarouco, and M. Bercht, "Initial Perception of Virtual World Users: A Study about Impacts of Learning Styles and Digital Experience", International Journal for Innovation Education and Research, vol. 5(5), pp. 95-112, 2017.

[33] A. Balderas, A. Berns, M. Palomo-Duarte, J. M. Dodero, and I. Ruiz-Rube, "Retrieving Objective Indicators from Student Logs in Virtual Worlds", Journal of Information Technology Research (JITR), vol. 10(3), pp. 69-83, 2017.

[34] I. Duncan, A. Miller, and S. Jiang, "A taxonomy of virtual worlds usage in education", British Journal of Educational Technology, vol. 43(6), pp. 949-964, 2012, doi: https://doi.org/10.1111/j.1467-8535.2011.01263.x.

Reconstruction and Web-based Editing of 3D Objects from Photo and Video Footage for Ambient Learning Spaces

David Bouck-Standen, Alexander Ohlei, Sven Höffler, Viktor Daibert, Thomas Winkler, and Michael Herczeg
Institute for Multimedia and Interactive Systems
University of Luebeck
Luebeck, Germany
email: [bouck-standen, ohlei, daibert, winkler, herczeg]@imis.uni-luebeck.de, sven.hoeffler@student.uni-luebeck.de

***Abstract*—In ambient and mobile learning contexts, 3D renderings create higher states of immersion compared to still images or video. To cope with the considerable effort to create 3D objects from images, with the NEMO Converter 3D we present a technical approach to automatically reconstruct 3D objects from semantically annotated media, such as photos and more importantly video footage, in an automated background process. Although the 3D objects are rendered in a quality acceptable for the scenario presented in this article, they still contain unwanted surroundings or artifacts and will not be positioned well for, e.g., augmented reality applications. To address this matter, with 3DEdit we present a web-based solution allowing users to enhance these 3D objects. We present a technical overview and reference pedagogical background of our research project Ambient Learning Spaces, in which both the NEMO Converter 3D and 3DEdit have been developed. We also describe a real usage scenario, starting by creating and collecting media using the Mobile Learning Exploration System, a mobile application from the application family of Ambient Learning Spaces. With InfoGrid, a mobile augmented reality application, users can experience the previously generated 3D objects placed and aligned into real world scenes. All systems and applications of Ambient Learning Spaces interconnect through the NEMO-Framework (Network Environment for Multimedia Objects). This technical platform features contextualized access and retrieval of media.**

Keywords—Mobile media; Mobile learning; Ambient Learning Spaces; 3D Conversion; 3D Editing.

I. Introduction

Today, in our networked society people are living within their digitally enriched environments. Together with ambient and mobile technology, these various individual interconnections between physical and digital worlds play an important role. Thus, being and acting in the physical world is accompanied by the creation of technology-assisted environments, like through the creation and visualization of 3D objects [1].

Contemporary pedagogical approaches follow the assumption that humans learn individually and during all of their life. In our research context, one important goal is to offer these ubiquitous learning environments; we called them *Ambient Learning Spaces (ALS)*, as described by Winkler et al. [2]. The relatedness of body and space supports the individuality of the learning process. Together with the loss of spatial distance and the exponentially growing quantity of information, this induces new technical requirements, as a single individual is no longer capable of consuming and structuring the globally and permanently available information in its entireness [3]-[5]. ALS enables learners to structure information themselves using ambient technology in web-based applications like in mobile contexts on their smartphones. This seems to be fostering the construction of sustainable and mindful knowledge. In this setting, 3D renderings empower imagination, creativity and learning compared to still images or video [6]. In our research project funded for more than seven years by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), ALS consist of a mixed reality where body and space are extended by digitally artifacts. These are represented by peripheral, tangible, mobile, and wearable media [2] and are illustrated in Figure 1, which shows the proximity relationship of each class of media to a learner in ALS.

In ALS, media in general and especially mobile media become the carrier of information utilized in various contexts [7]. This supports the following learning objective: the learner creates contextualized and personalized media stored as digital data, which can be enriched by digital properties. Together, media and their digital properties form will be called *enriched media.*

For the backend of ALS, we have developed the *Network Environment for Multimedia Objects (NEMO)*, which is used

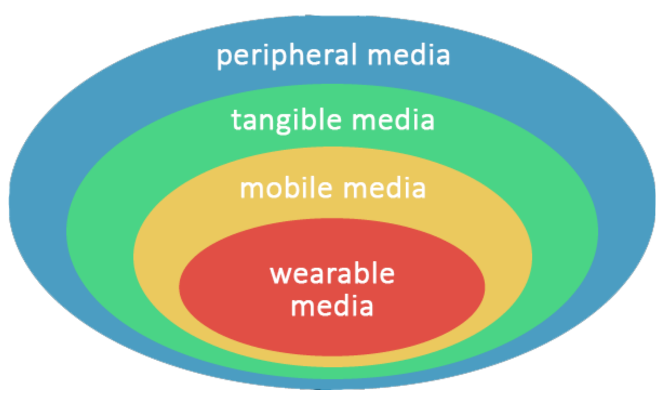


Figure 1. ALS shell model of media visualizing the proximity relationship between media classes and the learner [2]. The system concept of ALS is based on this structure.

as technical foundation of ALS in its latest implementation [8] based on the original concepts of Lob et al. [9]. NEMO is a web-based framework serving, among other capabilities, as a repository for enriched media. For this purpose, NEMO stores media such as text, still images, video, 3D objects, animations, and audio. Digital properties are stored in dynamic sets of semantic annotations and extend these media [8]. Apart from serving as a media repository, NEMO provides computational logic through a service-based interface for other applications developed for ALS. Using NEMO together with these *ALS Applications* provides a digital overlay for physical objects using enriched media within ALS.

Together with ALS applications, the NEMO framework stores media from users as enriched media, as outlined. These enriched media also contain images or video footage, which often originate from the same physical objects, only differing in visual angle, lightning or framing. With the *NEMO Converter 3D (NOC3D)* this paper presents a solution to make use of such footage collectively for example created by school students during field trips, converting these media into single 3D objects automatically without the need of user supervision. Additionally, the learner's experience of digital 3D worlds can further be enhanced by using the 3D Editor for ALS (3DEdit) making the use of a 3D editing program on a dedicated graphics computer system obsolete.

In this contribution, in Section II, we regard related work. In Section III, we present a real scenario using a smartphone application from the ALS application family. In Section IV, we introduce these ALS applications in more detail. The NEMO-Framework is outlined in more detail in Section V. In Section VI, we focus on the realization of NOC3D and describe 3DEdit in more detail in Section VII. In Section VIII, we present our findings and conclude with a summary and outlook in Section IX.

II. Related Work

Semantic media comprises the integration of data, information and knowledge. This relates to the Semantic Web [10] and aims at allowing computer systems as well as humans to make sense of data found on the web. This research field is of core interest since it yields naturally structured data about the world in a well-defined, reusable, and contextualized manner.

The field of metadata-driven digital media repositories is related to this work [11] as well. Apart from the goals of delivering improved search results with the help of meta information or even a semantic schemata, the NEMO framework distinguishes itself from a mere repository by containing and using repositories as internal components, delivering more sophisticated features through the NEMO logic described below.

NEMO facilitates collecting, consuming and structuring information by interacting device-independently with enriched media, whereas the linked data research targets sharing and connecting data, information and knowledge on the Web [12].

Various implementations exist in order to reconstruct 3D objects from photographic images. Those we have examined in our work have in common that they are not integrated into a fully automated web-based framework making use of semantically annotated data in mobile contexts providing background services for ambient learning environments.

In addition, various implementations exist in order to edit 3D objects, also in web-based applications. However, the implementations examined provide features for advanced and professional users and are directed at creating and editing 3D objects. Performing manipulations required for our scenario would require many complex steps of interaction and would not allow editing 3D objects in a touch-only application as needed in our learning scenarios.

In the research field of e-learning, other work connecting semantic structures with learning can be found [13]-[15]. In contrast, our work focuses on linking educational contents with the living environment (Lebenswelt) and thus engaging learners in communicative processes through contextualized and personalized enriched media. For this purpose, NEMO provides means of connecting formal and non-formal learning inside schools or outside of schools like in museums. NEMO is not used to examine or track the learner's performance, provide standard learning materials or collect homework, such as Moodle [16].

III. Scenario

Michelle, a fourteen year old student, joins a field trip through the Hanseatic City of Luebeck at school. Prior to the excursion, Michelle's teacher prepared some exercises for the students with the help of the ALS-Portal. They have to answer questions like "*What is communication?*" using the ALS application MoLES. While exploring the city, Michelle answers this question with MoLES running on her smartphone. Michelle uses MoLES and takes photos and tapes videos of what she thinks is related to the question at hand. In this case, for instance, she discovers a sculpture of four adults, two standing next to each other, and two sitting on a bench. From their body language, it seems that they are talking to each other. Michelle uses MoLES on her smartphone to take a few photos and to record videos. For every medium she creates, Michelle also takes notes using MoLES in form of some keywords and sentences to remember her thoughts later on. MoLES uploads these enriched media automatically into NEMO over a secure connection.

Back in school, Michelle prepares a short presentation of her findings from the field trip. In the meantime, Michelle found out, that the sculpture she took photos of is called "Neighbors in Conversation" and belongs to a series of artistic work. She logs on to the InteractiveWall located in the foyer at school. She browses through her media using swipe gestures on the multi-touch screen. Among her media, she discovers that the sculpture she took photos of is meanwhile available as an automatically created 3D object. Using the ALS InteractiveWall, she views the 3D object in full screen mode and notices that the sculpture is shown from its backside. She rotates the 3D object using swipe gestures until she is satisfied with the position. Looking at the sculpture from a bird' eye view, she notices that there are disturbing artifacts around the sculpture. Using 3DEdit, which is embedded into the InteractiveWall, she cleans up the model

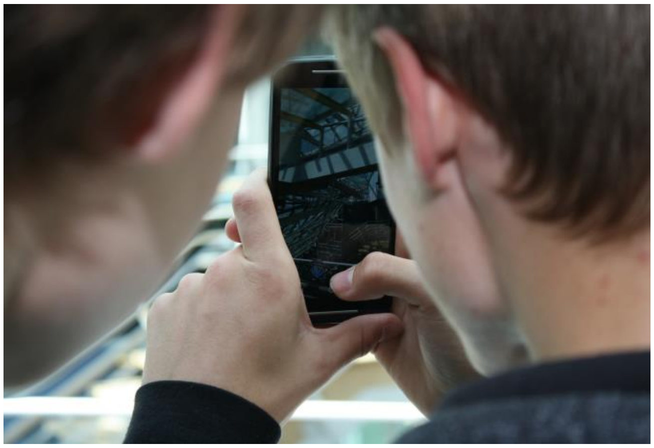


Figure 2. School students use MoLES to take photos and record videos to answer questions during a field trip.

by selecting only the sculpture and thereby removes the artifacts. After finishing, she is happy with the orientation and presentation of the rendering.

Now, Michelle incorporates the 3D object into her presentation. During the presentation, some of her classmates are surprised that they did not notice the sculpture themselves before.

With the help of the mobile application InfoGrid, they are now able to take a closer look at it from all sides. They are astonished to hear from Michelle's presentation on what she thinks is associated with the topic at hand.

Sometime later, Michelle is engaged in another school project and logs on to the ALS Portal. She browses through her media to find something suitable to use for the new project. She again comes across the sculpture of the "Neighbors in Conversation" and notices that a woman sitting on a bench is holding her child. She uses the 3D editor module inside the ALS Portal she used before on the InteractiveWall to cut out only the child-holding woman she then uses in her new project. She is satisfied that she can use the media again in other contexts.

IV. ALS Applications

In the following section, we are describing the frontend applications in ALS already mentioned and referred to in the scenario in more detail.

A. Mobile Learning Exploration System

The *Mobile Learning Exploration System (MoLES)* is a mobile ALS application running on smartphones and was originally introduced by Winkler et al. [17]. In a mobile context, students create enriched media to answer given questions for a specific task assigned by their teacher whilst conducting an exploration outside of school. They take photos

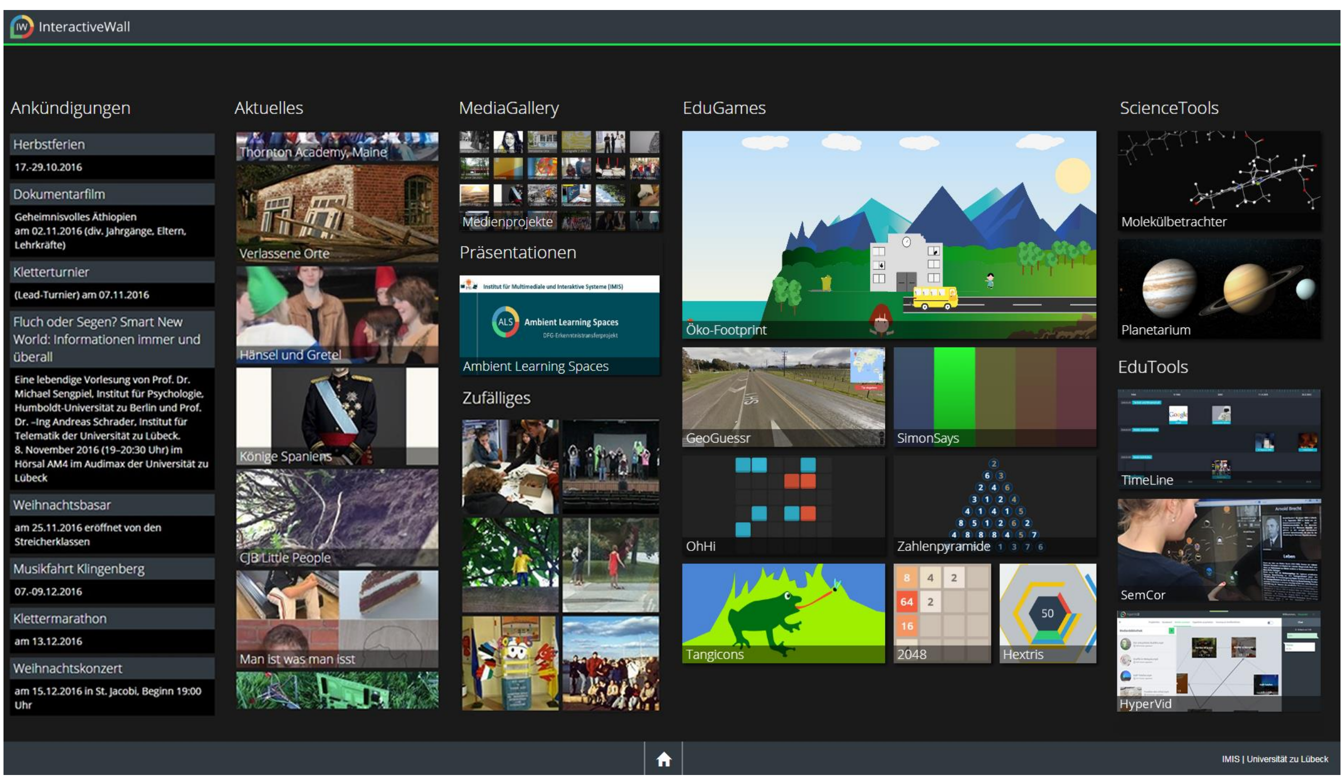


Figure 3. Start screen of the InteractiveWall, taken from one of our project partners. Apart from learning games and educational tools, the InteractiveWall visualizes media created by school students. These are shown in sections 'Latest' as well as 'Random' directly on the schools InteractiveWall. MediaGallery gives access to an overview of the entire media school students have access to.

and record audio or video footage from objects they encounter and add textual notes within MoLES by annotating the media, as illustrated in Figure 2.

After finishing a field trip, the students use the enriched media they created to reflect and present their findings to their fellow students. All media created with MoLES is stored in NEMO.

B. *InteractiveWall*

The *InteractiveWall* features an installation consisting of large wall-mounted multi-touch displays [7]. Students use the InteractiveWall to browse their digital media, the media created by others, and present their finding to their fellow students. The InteractiveWall features various applications in context of informal learning in schools, as Figure 3 of the starting monitor depicts and the users interact with the InteractiveWall through touch gestures only.

In our scenario presented in this contribution, Michelle uses the *MediaGallery* to browse the digital media she recorded during the field trip. On the InteractiveWall, among her media she discovers the sculpture she took pictures and recorded video of automatically reconstructed as 3D object by NOC3D and she uses 3DEdit to enhance the 3D object.

The InteractiveWall connects to NEMO. Thus, NEMO serves all media displayed on the InteractiveWall.

C. *InfoGrid*

InfoGrid is an augmented reality application for smartphones used in mobile context in ALS [18]. InfoGrid recognizes visual markers and detects Bluetooth beacons both triggering the display of images and video, the playback of audio or the augmented reality presentation and alignment of 3D objects. Figure 4 illustrates the display of InfoGrid.

Figure 4. Augmented Reality display with InfoGrid. On the left, the sculture "Birds" by Günter Grass and on the right, the skeleton of a blue whale as 3D object is shown.

The use of InfoGrid ranges from basic scenarios where markers are augmented by static media to more complex scenarios, where NEMO selects enriched media from its repositories with InfoGrid guiding the users on a dynamic narrative path.

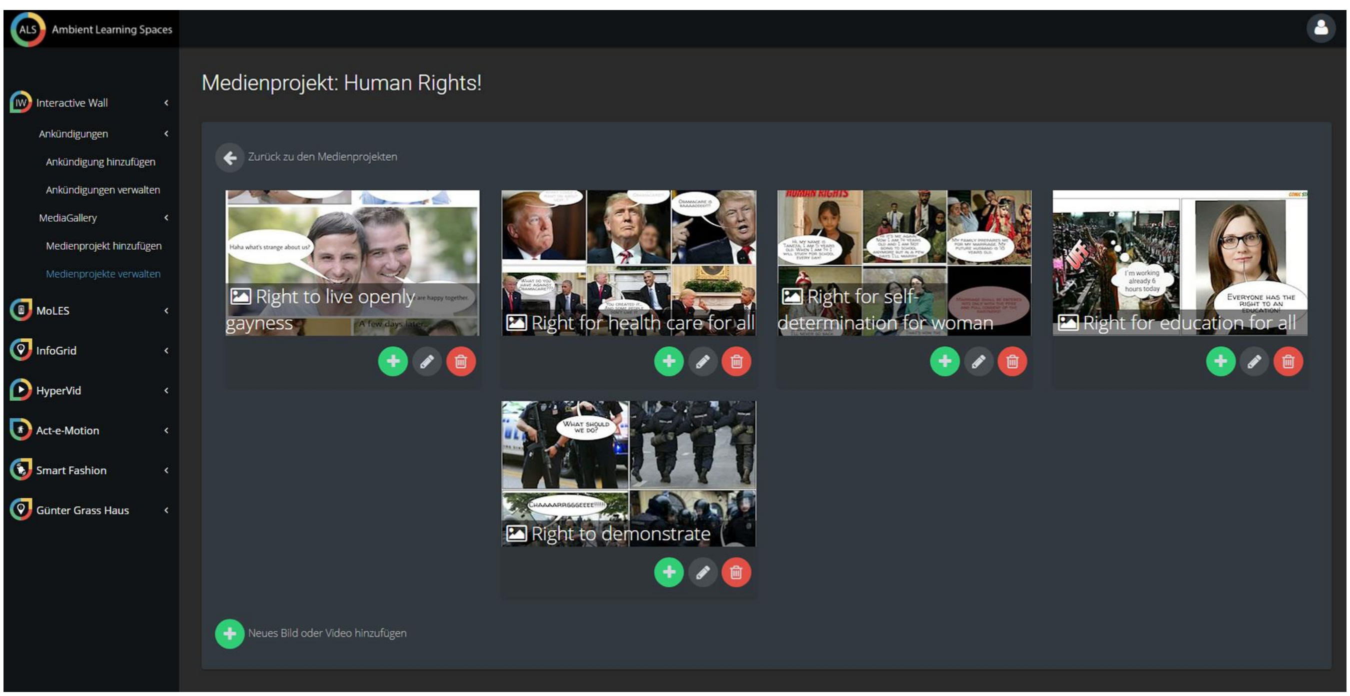


Figure 5. Screenshot taken from the ALS Portal of a school. The left menu offers a module selection. The current view shows an overview of a media gallery for the InteractiveWall. The media can be created, arranged, edited and deleted from this view by teachers or students.

In our scenario, school students use a single marker to view 3D objects automatically reconstructed by NOC3D in their classroom. For this, NEMO maps multiple media to this marker and delivers them to InfoGrid.

D. *ALS Portal*

The *ALS Portal* features the management of enriched media [6]. The ALS Portal is a modularized web-based platform. For each ALS application, a dedicated module within the ALS Portal allows to manage these enriched media, depending on the user's access rights and permissions.

As illustrated in Figure 5, the ALS Portal allows editing, e.g., a media gallery within the module MediaGallery. Media entered here is stored in NEMO together with semantic annotations an also made available by NEMO in other contexts of ALS.

In our scenario, Michelle is able to browse the media recorded with MoLES inside the ALS Portal. She also uses the editor 3DEdit to enhance the 3D object of the sculpture of the "Neighbors in Conversation".

All ALS applications interconnect through NEMO. This means, that enriched media created with MoLES, is also available on the InteractiveWall, can be edited using the ALS Portal.

The ALS Portal, InteractiveWall, and MoLES are web-based and implemented as ASP.NET applications. They connect to NEMO via Web Services.

V. NEMO

NEMO is a web-based framework for ALS. As depicted in Figure 6, the framework primarily consists of three main layers: (1) the NEMO Application Programming Interface (API) layer giving ALS applications access to NEMO, (2) the NEMO Logic layer and (3) the NEMO Core layer. NEMO as well as NOC3D have been implemented in C# running on Windows Server and Microsoft .NET architecture, also making use of the Windows Communication Foundation (WCF) framework.

The NEMO API (cf. Figure 6) provides access for applications such as MoLES, interacting through Web services in an authenticated context over a secure connection. Each application accesses a specific Web service, which achieves a higher layer of transparency and maintainability with regard to the system's architecture. With the NEMO API Client Model, we created a model for a well-defined data interchange between NEMO and any application in ALS through the Web, following the idea of knowledge representation in a formal and explicit way. From experience, we expect any information entered by a learner to be incomplete, as he or she is still engaged in a process of gathering, structuring, and memorizing, thus NEMO is able to handle incomplete and uncertain information [19] in the NEMO Logic and Core layers. Therefore, the model for any client application is independent of any internal model used by the NEMO framework. In addition, this minimizes the learning curve for ALS application development, as no detailed knowledge of semantic modeling is required when developing an application accessing NEMO. NEMO also provides cross-device capabilities [20].

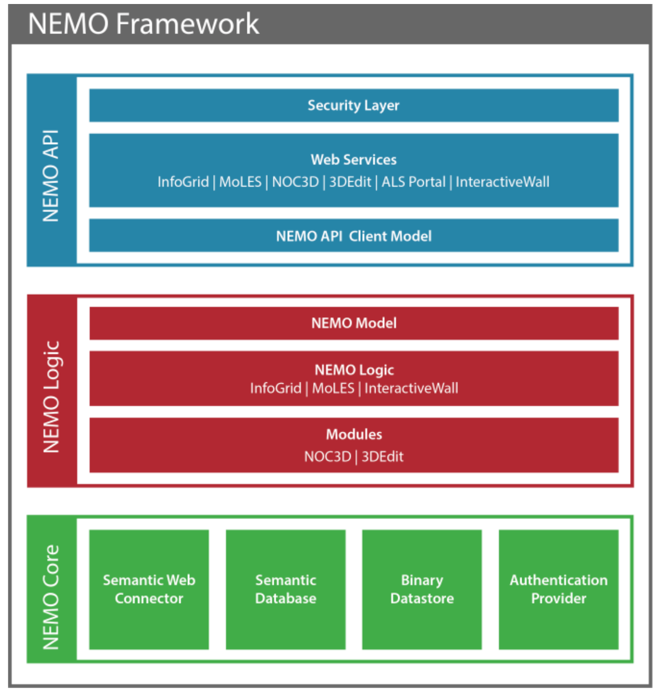


Figure 6. The NEMO-Framework [1]. The NEMO Logic computes, e.g., coherences, semantic models, and data mapping and a modularised interface for feature extendibility.

In the NEMO Logic, we implemented the NEMO Model, which abstracts ALS as a semantic model. Here, the computational logic resides. It initiates and controls semantic search and context analysis in the NEMO Core. In the NEMO Logic, mappings are conducted between the NEMO Model and the NEMO API Client Model through a modified Semantic Object Relational Mapping (SORM). For any Web service, the NEMO Logic holds the specific application logic and thus interconnecting the applications accessing the NEMO framework semantically through an extendable modular structure with loose coupling. We have already developed extensions for NEMO, e.g., the NEMO Converter (cf. Figure 6), which delivers media in device-specific formats and resolution as requested. For research purposes, another extension tracks all requests, actions, as well the corresponding application state of the NEMO framework. NOC3D also extends NEMO Logic. All data collected is stored anonymously due to the sensitivity of the data and legal regulations for public organizations like schools and museums. In a defined context of an evaluation, personal information may be collected synonymously. As we develop NEMO with technical scalability and diversity in mind, NEMO also runs in multiple interconnected instances.

The NEMO framework is based on the NEMO Core where enriched media is stored (cf. Figure 6). A semantic database provides internal storage for any digital entity in the form of semantic annotations. Through the Semantic Web Connector, (cf. Figure 6) any semantic database can be used as internal data store, thus developing applications accessing NEMO requires no knowledge of the respective database query language. Any query result of the internal or any external

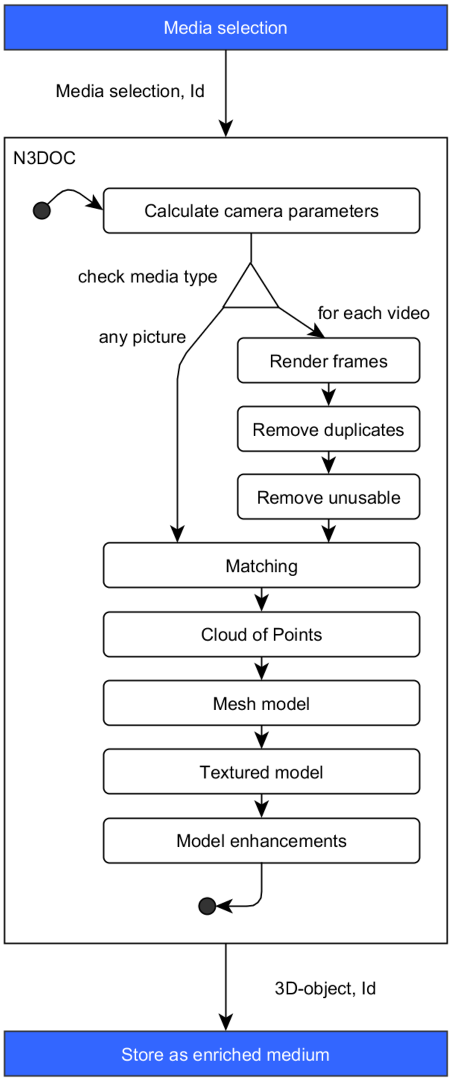


Figure 7. Pipeline of the NOC3D algorithm, as more detailed described in section V.A. Media selection as well as storing the 3D object is performed externally from NOC3D by NEMO.

semantic database is mapped into the NEMO Model. In the NEMO Logic, this data will be processed as described above. Binary media is stored in the Binary Storage (cf. Figure 6), which is linked to the internal semantic database in order to retrieve the stored object as enriched media again and also serves as cache in order to reduce on-the-fly conversion time of the NEMO Converter.

An authentication module provides an interface to connect to different authoritative systems in order to check application or media-specific permission settings and user access rights.

VI. The NEMO Converter 3D

NOC3D was developed as a component for the NEMO Logic under the following assumptions, which are partly derived from the scenario described above:

Figure 8. Screenshot of a 3D object reconstructed with NOC3D from 225 images automatically extracted from semantically annotated videos. The blue background is rendered by the 3D object viewer.

- NOC3D runs in an autonomous mode as a background service without any user interaction required.
- Images and videos are taken with different camera models, mostly with smartphones, from various angles and may contain only sections of the object. Therefore, an input for NOC3D can most certainly not be described as “ideal” or “complete”. The cameras have not been calibrated.
- The process of media creation does not require additional markers, only steady surroundings around the object. Every photo or video has to contain surroundings around the object, which is subject to reconstruction.
- An object for reconstruction has dimensions between 5cm and 5m in both width and height.
- Images and videos may not contain multiple objects and only one object will be reconstructed per run.

A. 3D Reconstruction

In general, the algorithms used in each step and data they require or provide as input and output determine the sequence of steps of the 3D reconstruction process. For our scenario in an ambient context, we have enhanced their combination and derived parameters from the tests we conducted.

At first, from manually entered and automatically generated semantic annotations, such as GPS coordinates, date and time and with regard to different calendrical seasons, NEMO compiles a selection of images and videos, which possibly show the same object which might be reconstructed.

Using GPS coordinates, date and time as well as data from Exif information [21], such as camera make and model, enhances the picture selection in NEMO.

All media selected by NEMO is transferred to the NOC3D module, as shown in Figure 7. As NOC3D provides a web-based API, NOC3D may be set up on a dedicated server, still being part of NEMO [8]. An identifier passed additionally allows NEMO to link the original media with the 3D object after the asynchronous task of NOC3D completes.

Operating on the media selection passed on by NEMO, NOC3D at first calculates camera parameters, which will be used for the process of reconstruction later on. 3D object reconstruction starts by calculating match points of all images and grouping them using VisualSFM [22]. This is necessary

Figure 9. Example of the media taken as still images or extracted from video footage Michelle took during her field trip. Input images for NOC3D are not expected to be ideal or oriented.

in order to find the object for 3D reconstruction in necessary in order to find the object for 3D reconstruction in the images automatically. Every two images with at least 40 match points are grouped. To receive a high quality result from later steps, all images with a resolution below 1200x1200px are discarded at this point. A group with less than 10 images is discarded as well, because these will not be of any use for further processing. We found these parameter values through experimental testing during development. Running VisualSFM on the group of images, until no other image of the selection can be grouped repeating all steps outputs a group of images containing the object for 3D reconstruction.

In the next step, depicted in Figure 7, the Center for Machine Perception Multi-view Reconstruction Software (CMPMVS) [23] calculates the cloud of points using the camera parameters from the first step [24][25]. CMPMVS transforms the point cloud into a mesh model and separately calculates a preliminary texture.

The textured model is handed over to MeshLab [26]. Small artifacts are removed. In addition, MeshLab is used to close polygon gaps in the reconstructed model, remove devious edges and smooth the entire model. The result of an exemplary 3D object is illustrated in Figure 8.

For web-based, mobile device, and browser compatibility, in this stage an additional 3D object is created, in which the number of polygons is reduced to 30.000.

After conversion into a NEMO-compatible file format, NOC3D hands over the completed 3D object to NEMO. NEMO stores the 3D object together with the semantic annotations of the images used to reconstruct the model. The user may have to adjust these annotations, depending on the variety of annotations of the original media. Afterwards, the 3D object is available in NEMO. From here, it can now be retrieved by ALS applications, like the ALS Portal for further editing, or by the presentation applications such as InfoGrid and the InteractiveWall, to view the object in various contexts, as the scenario outlines.

B. 3D Reconstruction from Video Footage

In general, in the process of 3D reconstruction more images from different angles lead to qualitatively better results. During the development of NOC3D through qualitative evaluation with university students, we found out that taking hundreds of images (cf. Figure 9) of the same object does not integrate well with our usage scenario. In case of an entire class of 20 or more students, who take at least five images of the same object, NOC3D produces acceptable results. However, the challenge of acquiring sufficient footage with low effort remains.

The process of acquiring footage used for 3D reconstruction is simplified by supporting videos as input format. Assuming a video generally consists of at least 24 frames per second, just moving around the object taping a video will produce a sufficient amount of material. Before starting the process of reconstruction, the videos have to be pre-processed, as illustrated in Figure 7. The video frames are extracted frame-by-frame into images using FFmpeg [27] and stored temporarily. This leads to duplicate or similar images,

Figure 10. 3D object obtained from automated reconstruction from 176 images through NOC3D. Figure 8 shows exemplary images used in the reconstruction process for this 3D object. This is the default unmodified view, which appears offset. The black background is rendered by the 3D object viewer, in this case 3DEdit.

e.g., when the camera movement around the object is slow. These images do not contribute usable data for the object reconstruction process.

The solution is to remove all duplicates during pre-processing using the imaging library ImageMagick [28]. In addition, unusable images like from overexposed or black frames will be removed.

After pre-processing, all images extracted from the video footage are joined with other images for reconstruction. Our tests indicate, that at least one image (e.g., photo) not taken from a video is required in order to produce acceptable results. The reason are camera parameters stored in Exif data, which are usually not separately stored with each video frame. These are required for the process of 3D reconstruction by CMPMVS. As for our scenario, smartphones used to take photos and tape videos available today produce video footage in similar quality to pictures, which are sufficient for NOC3D, as illustrated in Figure 9.

C. *Running Time Issues*

With regard to 3D reconstruction, running time of the module is critical. Preparing the media for processing is performed with linear effort, including extracting usable still images from video as shown in Figure 7. All further steps require significantly more effort, depending on the number of images, the objects complexity and the image resolution [1].

During development, we found that integrating NOC3D directly on the same server with NEMO is unpractical, as 3D reconstruction in general results in high processor (CPU) utilization. Besides, 3D reconstruction performs faster on Graphics Processing Units (GPU) than on CPUs [29].

The solution we implemented is to run NOC3D on a dedicated server. Therefore, we extended NOC3D to connect with NEMO through over Web Services. As a result, prior to reconstruction NEMO transfers all footage to NOC3D, which stores all data temporarily on that server system. The process of 3D reconstruction is started after all footage has been transferred and NOC3D signals NEMO the completion of a conversion process via callback. Because we are using a dedicated server, we are now able to choose CMPMVS as Cloud of Points algorithm, which only runs on CUDA-enabled (Compute Unified Device Architecture) GPUs.

At this point, we did not further quantify running time or have performed tests with various GPU systems, because running time largely depends on the CMPMVS algorithms. However, the optimizations, like pre-processing images described above and earlier [1] can improve running time on all systems.

3DEdit is implemented in JavaScript, also using HTML and CSS, running inside the client's browser. 3D objects are reduced to 30.000 polygons for viewing purposes. Only the reduced 3D objects are transferred to the client for display in 3DEdit. Their size usually does not exceed 4MB and transfer time depends on bandwidth. On the server-side, only the full-size 3D objects are used in the process of editing. Thus, 3DEdit performs well on current devices. As for the algorithms, with which the full-size 3D objects are manipulated on the server-side, running time performs with linear effort.

D. *Quality of Models from Automated 3D Reconstruction*

As Figure 10 illustrates, the quality of 3D objects obtained from the process of automated 3D reconstruction is sufficient for usage with other ALS applications.

However, NOC3D is not capable of automatically placing a 3D object. This is required, if the 3D object was used with, e.g., InfoGrid. As InfoGrid and its augmented reality display places the 3D object on a photographic marker, it is required orienting the object first in order to display it placed correctly on that marker.

In addition, 3D objects might contain unwanted surroundings and artifacts (cf. Figure 11). As the currently implemented process of 3D reconstruction requires surroundings to be part of the original footage in order to work, isolating the physical object is not an option. This observation corresponds to our scenario, as images and video footage are not intentionally taken for the process of 3D reconstruction (cf. Figure 9). This requires tools of manually retouching the automatically generated 3D object, as our scenario illustrates.

Figure 11. 3D object rotated manually from default view. The 3D object presents some artifacts as unwanted surroundings. The blue background is rendered by the 3D object viewer, in this case MeshLab.

Figure 12. Screenshot of the 3D object in reorentation mode. The users can rotate the 3D object, until they are satisfied. Afterwards 3DEdit saves the changes to a copy, keeping the original 3D object.

VII. WEB-BASED EDITING OF 3D OBJECTS

As NOC3D is not capable of determining which specific parts of the footage are relevant for the user, or the actual physical object, the reconstructed 3D objects may include superfluous parts of the physical object's environment, such as the surrounding ground surface. Similarly, the desired focus of the object cannot be extracted reliably from the information available, e.g., from images or video footage, Exif tags, or semantic annotations. Thus, when viewed in their default orientation, the reconstructed 3D object may appear offset to the side or turned away from the viewer, as illustrated in Figure 10.

For this reason, with 3DEdit we have developed an additional web-based application, which offers two specific functionalities sufficient to solve these issues for our scenario:

(1) A function to reorient the 3D objects, which sets the object's center and default orientation according to the user's requirements.

(2) A function to cut extraneous parts of the 3D object.

To allow for editing in a mobile and ambient context, 3DEdit offers a browser-based interface that seamlessly integrates with the ALS Portal or the InteractiveWall and can be used on mobile devices. Through this interface connected to NEMO, the necessary functionalities are simplified and automated to the point where they only require a single input by the user. This way, even an inexperienced user can make the necessary adjustments easily without much effort.

The web-based editor component of 3DEdit makes use of JavaScript in order to display the 3D object. The controls are minimized to the functions presented to users. For the purpose of using 3DEdit as module for both the InteractiveWall and the ALS Portal, 3DEdit uses the CSS style sheets corresponding to the application it is used in. For this reason, some figures in this article show blue controls on purpose.

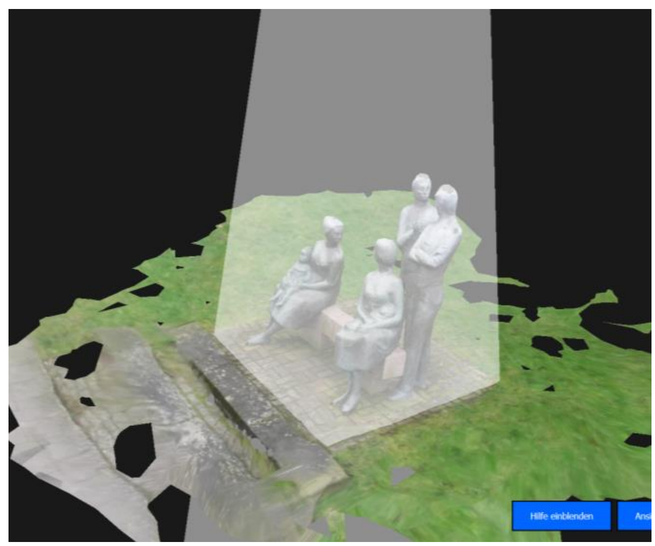


Figure 13. Screenshot of the 3D object in cutting mode. Extraneous parts of the 3D object are outside of the selected volume and not kept after applying the changes using 3DEdit.

Please also note that all screenshots of 3DEdit show 3D objects reduced to 30.000 polygons because of browser limitations. The actual 3D object is kept in NEMO in full size.

A. *3D Object Manipulation*

The actual manipulation of the 3D object is conducted inside NEMO, so there is no special hardware required in addition to the requirements of the ALS Portal and the InteractiveWall. The modules communication is handled via Web Services, which connects the frontend of 3DEdit to its dedicated backend component, which is located in the NEMO

Figure 14. Screenshot of the 3D object after reorientation and cutting is finished using 3DEdit from a web browser.

Figure 15. Screenshot of the 3D object in cutting mode. Also parts of a larger 3D object may be cut out using 3DEdit and use, e.g., in InfoGrid.

Logic layer, as outlined above. Running as part of the NEMO Logic, 3DEdit automatically makes use of the 3D platform *Blender*. This offers a consistent running time as well as a static location for depositing and retrieval of the media by ALS applications through NEMO.

The process of editing a 3D object is illustrated in Figures 12-14. In order to orient the misaligned 3D object (cf. Figure 10), the view is rotated until a suitable angle is found, as Figure 12 illustrates. This angle depends on the use case, in which the 3D object is used. Through experiments, we found that rotating the object toward a head-on view is sufficient for our scenario.

All 3D objects reconstructed from footage from physical objects from our scenario such as, e.g., statues and sculptures have in common to be standing on a base level and their height is limited. Considering this, unwanted surroundings can be removed by placing a clipping volume around the area of the object to keep. Any polygon outside the volume will be omitted or cut along the selected edge. On our InteractiveWall, the user places the volume inside the viewport and modifies its size with the help of touch gestures, as Figure 13 illustrates.

After the 3D object is aligned and all unwanted surroundings and artifacts have been removed in the editor of 3DEdit, the backend of 3DEdit takes the necessary actions to calculate the resulting 3D object, illustrated in Figure 14. At first, the parameters from the user's selection in the web-based interface is transferred to the backend and translated into modification commands. A single user selection of a volume containing the 3D object requires multiple commands executed by 3DEdit using Blender sequentially, in order to manipulate the 3D object accordingly.

B. *Quality and Running Time*

Using 3DEdit on 3D objects does not affect the objects quality, as 3DEdit only deletes or clips polygons using Blender.

The original 3D object is not displayed in the browser-based editor of 3DEdit due to limitations. In the backend, within NEMO 3DEdit manipulates the actual 3D object.

The running time for both reorientation and cutting modifications is linear. This supports the use of 3DEdit in our scenario. 3Dedit may also be used in order to extract certain

Figure 16. Screenshot of the cut out part of the 3D object. This becomes available as separate 3D object in NEMO.

Figure 17. Statue called "Dorothea" by the people of the Hanseatic City of Luebeck.
Number of images used for 3D reconstruction, from left to right: 62, 110, 233, 327.

parts of a larger model, as the scenario describes. Figures 15-16 illustrate this process. In terms of running time, there is no difference between removing unwanted surroundings or extracting parts of an existing 3D object. This can alo be observed with 3D objects, which have not been generated through NOC3D, but modeled using professional tools.

VIII. FINDINGS

In summary, NOC3D produces 3D objects with an acceptable quality given the mobile and ambient context of our scenario in the open standard OBJ-file-format. Due to the automatic process, it is inevitable that 3D objects may contain some surroundings, or will be misaligned with regard to their orientation. With 3DEdit we developed a solution to address these imperfections. This provides a web-based user interface that allows editing of 3D objects in alignment and removal of unwanted surroundings, as well as a backend module, which manipulates 3D objects on the server-side within the backend framework NEMO. 3DEdit can be used with a touch interface or by mouse.

In order to integrate NOC3D in a timely manner as outlined in our scenario, most importantly a multi-GPU system consisting of multiple CUDA-compatible graphic boards is recommended. In addition, free RAM capacity of at least the size of the footage used for conversion as well as hard disk storage of at least ten times the size of the footage for temporary storage is advisable. NOC3D supports images and video footage from different cameras and in different resolutions.

We have taken footage from more than 30 different statues across the Hanseatic City of Luebeck, Germany, and compiled them into different selections according to semantic annotations using NEMO. The footage taken cannot be described as 'ideal', as we cared to take mostly snapshots, e.g., only showing parts of the objects or without optimal lightning that would be used when reconstructing 3D models in, for instance, a laboratory with a special 3D scanner. Thus, our tests reflect media expected to be created by students on a field trip, matching our scenario.

With regard to the usage of MoLES in our scenario, the task illustrated was simplified for this contribution. In a real scenario, the question and tasks are accompanied with pedagogical considerations, which lead to a set of questions for each task, as outlined by Winkler et al. [30].

The process of selecting pictures by semantic annotations is not trivial, as our tests show. It depends on the quality of semantic annotations. Especially with manual annotations, the quality varies. Our observations from school students as well as teachers usage of the ALS Portal show a tendency to repeat and generalize semantic annotations. In many cases, groups of media are annotated with the same semantic annotations. We are currently working on a solution in order to encourage users to use more diverse semantic annotations when creating media in ALS. In a first approach, we additionally save and display annotations automatically created by a computer vision library, which is automatically accessed for each still image uploaded by NEMO. These automatic annotations are displayed as suggestions. However, we still have to evaluate possible impact on the users interaction and the quality of the annotations with regard to their use when selecting pictures.

The image selection process is further enhanced for best results by also using data from Exif tags, which are included in of enriched media, providing data on e.g., camera make, model, date, time and geolocation. This data is also used in the reconstruction process.

Our evaluation shows that, in our scenario, an average minimum of 110 images is required in order to be able to recognize the resulting 3D object as such, as illustrated in Figure 17. The maximum of images is limited by hardware resources, but keeping in mind the time-consuming process of

Figure 18. On the left: Statue “Panther” in a botanic garden in the Hanseatic City of Luebeck. On the Right: The output from 175 photos is hardly recognizable as a panther.

3D reconstruction should be limited to a maximum of 450 images. This value is derived from our experiments in context with our scenario and is depending on the objects complexity, desired quality, hardware capabilities, and the usage scenario. Hence, our research does not focus on optimizing the algorithms employed for 3D reconstruction within NOC3D, but we recommend setting up NOC3D on a dedicated GPU render server.

Further research and tests revealed that the quality of 3D objects is enhanced if the actual object for reconstruction fills about two-third of the entire image or video frame. This is due to the nature of the reconstruction algorithm.

Using footage from symmetric objects especially in front of symmetric or repeating backgrounds often leads to unusable 3D objects, as the example in Figure 18 shows. Using more photos does not necessarily enhance the output.

Generally and as expected, higher resolution of footage as well as using more images results in more detailed 3D objects, but also consuming more time during reconstruction. Nevertheless, using MoLES in contexts like our scenario limits students to the use of smartphones, which is why our primary focus lies on generating acceptable 3D models from smartphone-generated footage.

During our tests, we found that in some cases NOC3D aborted due to a memory overflow. This occurs due to limited hardware resources, exhausted by huge amounts of input data. Apart from upgrading the hardware, our solution is to catch the exception and remove images with the highest and lowest resolution gradually, restarting the process. With this strategy we try to keep as much information on the object and as much high quality footage as possible. This strategy may be optimized.

In total, NOC3D generates all sample models without any unexpected result or malfunctioning. Processing the sample models depends on model complexity and the amount of data to process. Using multiple GPUs is highly recommended.

We also tested the accompanying editor 3DEdit with a variety of 3D objects generated by NOC3D. In every case, it was possible to edit the generated object successfully in a matter of minutes, resulting in cleaned-up and properly oriented objects suitable for further use in other ALS applications. We have also tested 3DEdit with a number of 3D objects created with professional 3D tools containing up to 2.9M polygons.

IX. Summary and Outlook

NOC3D is a module for NEMO that serves fully automated reconstruction of 3D objects from images and most importantly from video footage created in ambient and mobile context and is used for learning scenarios in ALS. Through NOC3D, enriched media collectively created by students using their smartphones in mobile context is converted into 3D objects. Using web technology, we integrate 3D objects seamlessly with applications from ALS which are used in mobile contexts and in context of learning with media. These 3D objects can be enhanced using two steps of the web-based 3D object editor 3DEdit, in order to reorient the 3D object, remove unwanted surroundings or cut out a certain part of the entire object.

It is our hypothesis that learning in a formal and non-formal learning space [7][30], which is digitally enriched through ambient media, fosters cognitive skills and knowledge in a communicative environment [31][32]. We are going to evaluate this in more detail in formal context within schools as well as non-formal context in museums.

3D objects play a vital role in our research. We plan to evaluate their values in a digitally enriched learning environment. In the setting of our ongoing research, InfoGrid will be deployed to our four project partners, two schools and two museums within this year. This will allow us to further develop our scenarios and to examine, what is gained from a learning perspective, by the integration of 3D objects into Ambient Learning Spaces.

For our school project partners, we are currently developing new teaching models implementing ALS technology.

For our museum project partners, we have developed scenarios integrating 3D objects exhibitions with special focus on the autonomous use inside the museums. In the museum context, NOC3D can be also used in the process of reconstruction among others for objects of cultural heritage. Although for this use case dedicated digitization systems exist, like using accurate laser scanning, in the case where only photographic or video material of an object remain for certain reasons, the use of NOC3D is imaginable. In museum contexts, with 3DEdit the curators experience during the process of selection of 3D objects to augment an exhibition can be enhanced, as 3DEdit offers a solution to enable curators to edit 3D objects from any source. From our project partners we learned that this might be the case, when only a special part of a 3D object should be exhibited. In our further work, we are going to evaluate the curators’ user experiences using or ALS systems.

Especially in museum contexts, 3DEdit will also be used together with InfoGrid in order to align 3D objects on photographic markers. This gives the curator’s a tool to make fine adjustments to the presentation of their digital museum contents.

For any ALS application of the research project, among other features, NEMO provides persistent semantic storage of enriched media. Together with our project partners, two schools and two museums located in the Hanseatic City of Luebeck, the use of these applications together with NEMO in context of mobile and ambient learning is currently being further evaluated. NEMO is running in multiple instances on-site. The ALS applications are featuring the creation, presentation, use and interaction with enriched media. The applications are developed for various platforms, in desktop, stationary and mobile contexts. Thus, NEMO and the ALS applications are connecting the learner's knowledge with an ambient context, bridging the learning environment and lived-in world (Lebenswelt) to foster sustainable learning and meaningful knowledge.

ACKNOWLEDGMENT

We develop NEMO in the research project "Ambient Learning Spaces" supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG). We thank our project partners for their support. We extend our thanks especially to the Museum for Nature and Environment and the Günter Grass-House, both located in Lübeck, to the Günter und Ute Grass Stiftung, and to the Steidl Verlag.

REFERENCES

[1] D. Bouck-Standen, A. Ohlei, V. Daibert, T. Winkler and M. Herczeg, "NEMO Converter 3D: Reconstruction of 3D Objects from Photo and Video Footage for Ambient Learning Spaces," *AMBIENT 2017 - The Seventh International Conference on Ambient Computing, Applications, Services and Technologies,* IARIA, pp. 6-11, 2017.

[2] T. Winkler, F. Scharf, C. Hahn and M. Herczeg, "Ambient Learning Spaces," *Education in a Technological World: Communicating Current and Emerging Research and Technological Efforts,* pp. 56-67, 2011.

[3] A. Lugmayr, T. Risse, B. Stockleben, K. Laurila and J. Kaario, "Semantic ambient media - an introduction," *Multimedia Tools and Applications*, vol. 44, no. 3, pp. 337-359, 2009.

[4] M. McLuhan, "Understanding Media: The Extensions of Man," McGraw-Hill, New York, 1964.

[5] A. Whitmore, A. Agarwal and L. Da Xu, "The Internet of Things - A survey of topics and trends," *Information Systems Frontiers,* vol. 17, no. 2, pp. 261-274, 2015.

[6] K. Persefoni and A. Tsinakos, "Use of Augmented Reality in terms of creativity in School learning," in *Make2Learn 2015 workshop at ICEC'15*, Trondheim, Norway, pp. 45-53, 2015.

[7] T. Winkler, D. Bouck-Standen, M. Ide, A. Ohlei and M. Herczeg, *InteractiveWall 3.1 - Formal and Non-Formal Learning at School with Web-3.0-based Technology in Front of Large Multi-touch Screens.* In Johnston, J.P. (Ed.) *EdMedia: World Conference on Educational Media and Technology.* Washington, *DC, USA: AACE, pp.* 1317–1326, 2017.

[8] D. Bouck-Standen, "Construction of an API connecting the Network Environment for Multimedia Objects with Ambient Learning Spaces," Master Thesis, Lübeck, Germany, doi: 10.13140/RG.2.2.12155.00804, 2016.

[9] S. Lob, J. Cassens, M. Herczeg and J. Stoddart, "NEMO - The Network Environment for Multimedia Objects," *ACM (Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia (IITM 2010), Dec. 27-30, 2010, IIIT Allahabad, India),* pp. 245-249, 2010.

[10] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web," *Scientific American,* pp. 30-37, 2001.

[11] F. Nack, "The future in digital media computing is meta," *IEEE MultiMedia,* vol. 11, no. 2, pp. 10-13, 2004.

[12] C. Bizer, T. Heath and T. Berners-Lee, "Linked Data - The Story So Far," *International Journal of Semantic Web Information Systems,* vol. 5, no. 3, pp. 1-22, 2009.

[13] S. S. Kusumawardani, L. E. Nugroho, A. Susanto, A. Kumara, H. S. Wasisto and U. Cortés, "Ontology Development of Semantic E-Learning for Final Project Course," *Advanced Science Letters,* vol. 21, no. 1, pp. 46-51, 2015.

[14] M. Masud, "Collaborative e-learning systems using semantic data interoperability," *Computers in Human Behavior,* vol. 61, pp. 127-135, 2016.

[15] P. Bouquet and A. Molinari, "A New Approach to the Use of Semantic Technologies in E-Learning Platforms," *International Journal of Advanced Corporate Learning,* vol. 9, no. 2, pp. 5-12, 2016.

[16] M. Dougiamas and P. Taylor, "Moodle: Using Learning Communities to Create an Open Source Course Management System," *World Conference on Educational Multimedia, Hypermedia and Telecommunications,* pp. 171-178, 2003.

[17] T. Winkler, S. Günther and M. Herczeg, "Moles: Mobile Learning Exploration System," in *Proceedings of Society for Information Technology & Teacher Education International Conference (SITE)*. AACE. Charleston, SC, USA, pp. 3230-3234, 2009.

[18] A. Ohlei, D. Bouck-Standen, T. Winkler, J. Wittmer and M. Herczeg, "InfoGrid4Museum: A Medial Strategy for Mediation in Museums using Augmented Reality," *47. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Gesellschaft für Informatik (GI),* Available: http://hci.uni-konstanz.de/downloads/7_InfoGrid4Museums_Ohlei.pdf, 2017.

[19] P. Oliveira and P. Gomes, "Instance-based Probabilistic Reasoning in the Semantic Web," in *Proceedings of the 18th International Conference on World Wide Web*. ACM, New York, pp. 1067-1068, 2009.

[20] D. Bouck-Standen, M. Schwandt, T. Winkler and M. Herczeg, "ELBlocks - An Interactive Semantic Learning Platform for Tangibles," *Mensch und Computer 2016 – Workshopband*, doi: 10.18420/muc2016-ws10-0002, 2016.

[21] CIPA, "Exchangeable image file format for digital still cameras: Exif Version 2.3," in *Standard of the Camera & Imaging Products Association*, Camera & Imaging Products Association, 2012.

[22] C. Wu, "VisualSFM : A Visual Structure from Motion System," 2011. [Online]. Available: http://ccwu.me/vsfm/. [Accessed 20 May 2017].

[23] Center for Machine Perception, Czech Technical University, Prague, [Online]. Available: http://cmp.felk.cvut.cz. [Accessed 20 May 2017].

[24] M. Jancosek and T. Pajdla, *Multi-View Reconstruction Preserving Weakly-Supported Surfaces, CVPR 2011,* IEEE

Conference on Computer Vision and Pattern Recognition 2011, pp. 3121-3128, 2011.

[25] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multi-View Stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 8, pp. 1362-1376, 2010.

[26] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli and G. Ranzuglia, "MeshLab: an Open-Source Mesh Processing Tool," in *Eurographics Italian Chapter Conference*, pp. 129-136, 2008.

[27] F. Bellard, "FFmpeg," [Online]. Available: https://www.ffmpeg.org/. [Accessed 15 January 2018].

[28] ImageMagick Studio LLC, "ImageMagick," [Online]. Available: https://www.imagemagick.org. [Accessed 12 December 2017].

[29] M. Denkowski, "GPU accelerated 3D object reconstruction," *International Conference on Computational Science, ICCS 2013,* vol. 18, pp. 290-298, 2013.

[30] T. Winkler and M. Herczeg, "The Mobile Learning Exploration System (MoLES) in Semantically Modeled Ambient Learning Spaces," *IDC '13 Proceedings of the 12th International Conference on Interaction Design and Children,* pp. 348-351, 2013.

[31] C.-C. Huang, T.-K. Yeh, T.-Y. Li and C.-Y. Chang, "The Idea Storming Cube: Evaluating the Effects of Using Game and Computer Agent to Support Divergent Thinking," *Educational Technology & Society,* vol. 13, no. 4, pp. 180-191, 2010.

[32] M. D. Dickey, *Engaging by design: How engagement strategies in popular computer and video games can inform instructional design*, vol. 53, Kluwer Academic Publishers, pp. 67-83, 2005.

An Introduction to Edge Computing and A Real-Time Capable Server Architecture

Volkan Gezer, Jumyung Um, and Martin Ruskowski
Innovative Factory Systems (IFS)
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
Emails: {name.surname}@dfki.de

***Abstract*—The Internet has changed the way people access the information they need, and indeed how they live. Whether it is individuals reading emails or watching videos, or factories utilising automated fabrication devices, the access and processing of data is totally different. Thanks to the accessibility and the benefits that it brings into the lives, new research areas are emerging. One of the areas is Internet of Things (IoT) which connects countless of devices to the Internet. Increasing usage in IoT tremendously increases the count of connected devices to the Internet as well as the data generated and transferred. However, this increase brings several issues which could degrade the Quality of Service (QoS) with delays or even failed requests due to bandwidth limitations. Current tendency to solve problems that the Cloud Computing has is to perform computations close to the device as much as possible. This paradigm is called Edge Computing. There are several proposed architectures for the Edge Computing, but there is not an accepted standard by the community or the industry. Besides, there is not a common agreement on how Edge Computing architecture physically looks like. In this paper, we describe the Edge Computing, explain how its architecture seems, its requirements, and enablers. We also define an extensible, server architecture. The proposed Edge Server architecture has an ability to decide whether the task should be offloaded to the Cloud or to another Edge Server by considering the several parameters such as available resources and network delays. The resources of Edge Server can be extended with additional optional hardware or software modules to add new functionalities for artificial intelligence tasks, additional storage, wireless communication, etc. The server in the proposed architecture is also capable of performing real-time tasks and uses standard technologies to keep migration efforts at minimum. The paper also shows the results of an initial experiment, done without and with an Edge Server to compare computing performance.**

***Keywords*–Edge computing; real-time computing; edge computing requirements; enablers; Fog computing.**

I. INTRODUCTION

Internet of Things (IoT) gave new possibilities and changed how people live their lives. Number of connected devices to the Internet is going up with the increased tendency towards IoT [1]. In 1992, ”connected device” count was around one million which went up to 500 million in 2003 thanks to increased usage of personal computers. Later, IoT became even more popular and saw three billions of connected devices. In 2012, inclusion of of wearable devices increased this number to 8.7 billion. In 2013, this number went up to 11.2 billion owing to connected home appliances and in 2014, 14.4 billion with smart grids. The numbers increased in the upcoming years due to involvement of even small personal objects, such as toothbrushes, traffic lights, and table watches. Finally, even door levers are expected to be part of smart objects in 2020 [2].

Researches foresee that the connected devices are expected to be around 50 billion by 2020 [2][3]. This number is high as the Cyber-Physical Systems (CPS) and more intelligent components being used even for simple tasks. Tendency towards Cloud Computing and IoT devices leveraged the research in this domain and created new ones.

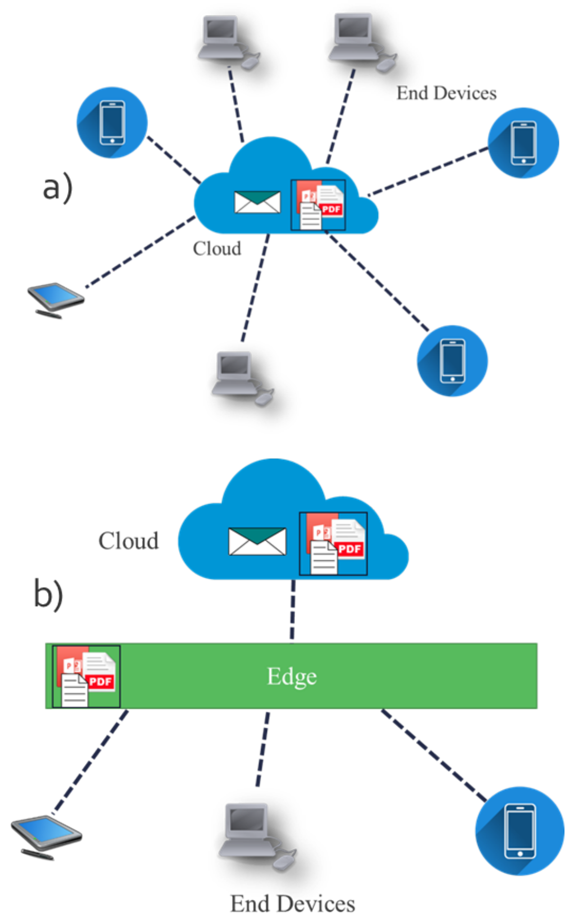


Figure 1. A simplified example showing the major difference between Cloud and Edge Computing. Cloud Computing (a) connects end-devices to the Cloud directly whereas Edge Computing (b) has an additional computing power in-between.

Cloud Computing or *the Cloud*, allows its users to store data, perform tasks using data centres through the Internet. The available resources in the Cloud granted low-powered or resource limited end-devices perform complex tasks in the Cloud, saving exceptional computational time [4]. Thanks to the ubiquitousness of the Cloud, data can be accessed from anywhere and any time as long as an active Internet connection is available. Some everyday tasks such as checking e-mails, video streaming, photo browsing, and file sharing or industrial tasks such as getting sensor values or controlling robots are performed through the Cloud. Using different standards, a single infrastructure to keep the system reliable is becoming even more complex, causing difficult and costly maintenance. The companies and research institutes are working to avoid failures of tasks due to insufficient hardware and network resources. The physical distance to the Cloud and the available resources within the infrastructure increase the latency and reduce the Quality of Service (QoS). One of the recent paradigms in this area to solve issues of Cloud Computing is Edge Computing. Although there are several naming for Edge Computing such as Fog Computing and Cloudlets, within this paper, only the term *Edge Computing* will be used. Figure 1 shows the difference between Cloud and Edge Computing.

Cloud Computing [5] is an emerging technology which allows machines/people to access the data ubiquitously. It enables on-demand sharing of available computing and storage resource among its users which could be either human or machine, or even both. Today, it is even possible for a simple device to share its status or get information over Internet with millions of users.

A layer is a logical organisation of set of services, devices, or software with the same/similar specific functionality, mainly defined for abstraction of tasks. A tier is, however, a physical deployment of layers for scalability, security and to balance performance [6]. In Cloud Computing, communication between a device and the infrastructure which provides the service is direct, without involvement of other tiers. In Edge Computing, however, an intermediate component, or an Edge Server performs the initial computation.

Edge Computing combines multiple technologies such as Cloud Computing, Grid Computing, and IoT. It adds an additional tier between the Cloud and the end-devices and moves computational power to the end-device as close as possible. This means that, in the need of more computational resource by the end-device or a system, the task can be offloaded to an Edge Server instead of the Cloud. Edge Computing is expected to reduce the latency and increase the QoS for tasks which cannot be handled by these devices. These tasks are usually computationally heavy such as big data processing, video processing, artificial intelligence or time-sensitive. If the computation must be done in real-time, utilization of Cloud is out of the question since Cloud and Internet offers only best-effort service and delivery. A system is a real-time system only if it reacts to its environment by performing the correct predefined actions within the specified time intervals.

Real-time computing can be divided into three categories:

- *Hard Real-Time*: Failure in the system is mostly fatal. For example, if an airbag in a car deflates before or after the specified timeframe (between 100 ms and 300 ms, within 10 ms), it loses its protective impact [7].
- *Firm Real-Time*: A real-time category between hard and soft real-time. It tolerates some deadline misses, but increase in the misses degrades the service, in the end causing unacceptable results [8]. For example, miss-sorting colors of the parts are acceptable up to some point [9].
- *Soft Real-Time*: This category groups the real-time applications which are less critical and have wider deadline interval for their acceptance. For example, voice calls or video streams are tolerated in case some data packages are lost.

Some systems produce gigabytes of data per second [10][11]. Devices with limited computing capacity may also have critical deadlines for their primary task. In these situations, the task can be offloaded to an Edge Server using the same constraints and can be accomplished at this level. Depending on the outcome of the task, the system reacts to the result, e.g., sends the data back to the end-device.

Both Edge Computing and Cloud Computing are strongly related to IoT and allow accessibility of the data ubiquitously. To build an architecture, the issues on the current Cloud or IoT systems must be identified, requirements must be specified, enabling technologies must be listed, and then a concept must be given. Later, the concept can be implemented in an architecture, validated, and evaluated. This paper presents an ongoing work on Edge Computing with its clear description. It also explains its requirements and enablers to solve the introduced issues. The paper also shows an ongoing work to implement a novel server architecture which is capable of performing real-time tasks and take the necessary actions to provide a high QoS, such as offloading the task to another server or to the Cloud. The architecture will not simply be another architecture, but will compare the existing architectures and consider real-world requirements from the industrial use cases. The architecture will also be vendor-independent and extensible and it meet industrial requirements.

The rest of the paper is structured as follows: Section II introduces some of the existing work done and simulators in the area of Edge Computing. Section III describes the concepts and some definitions together with requirements and enablers. Section IV defines the Edge Server architecture to be implemented for Edge Computing in two sub-sections. Section V shows the initial experiment results with various scenarios and finally, Section VI concludes the paper and presents the future work.

II. Background

Although usage of the term "Edge Computing" is recent, there are already several proposed architectures available, each considering different aspects to meet the requirements of the Edge Computing.

The architecture proposed by IBM considers the requirements for autonomy and self-sufficiency of production sites. The architecture is three-layered to balance the workload between the Edge, the Plant, and the Enterprise. The challenges of the architecture are listed as productivity gains for high throughput, failure prevention for reliable system and high product quality, and flexibility while hiding the complexity and allowing reconfiguration without a lot of effort [12].

Another reference architecture is proposed by OpenFog Consortium [13]. This architecture names the core principles as pillars. Pillars group requirements within their scope. These pillars are Security, Scalability, Openness, Autonomy, Agility, and Programmability. OpenFog Reference Architecture is proposed by covering industrial use cases.

Another recent initiative to build a common platform for Industrial IoT Edge Computing is EdgeX Foundry [14]. It was launched by Linux Foundation and initial contribution made by Dell. However, similar to OpenFog Consortium, it is also open for new memberships. EdgeX Foundry is a vendor-neutral open source software platform that interacts at the Edge of the network. It defines its requirements in architectural tenets as follows: platform agnostic in terms of hardware and operating system, flexible in terms of replaceability, augmentability, or scalability up and down, capable in storing or forwarding data, intelligent to deal with latency, bandwidth, and storage issues, secure, and easily manageable. A similar framework called Liota is being developed by VMware and it also aims at easy to use, install, and modify. Secondarily, it targets for a general, modular and enterprise-level quality. This framework is also open source and governed by VMware [15].

There are also several work done for computation and control in the Cloud. Below some of the related work is explained.

A research project called ”pICASSO” focuses on the control of a robot using a Cloud-based control platform. The project implemented a platform and a Cloud controller which can perform motion planning and control for industrial robots [16].

A recent work done by Givehchi, Imtiaz, Trsek, and Jasperneite [17] studies industrial use cases for using virtual control service in a private Cloud. Instead of using hardware programmable logic controllers (PLC) on site, they use a computer with multi-core processor and set each core as a virtual PLC to control sensors and actuators. The solution suggests a low-cost, but a slightly lower performance software PLC, compared to the hardware PLCs.

Another study on Cloud-based control is done by Goldschmidt et. al [9]. The work introduces a new architecture for scalable and multi-tenant Cloud-based control, virtualized PLCs. It also considers and evaluates the architecture with respect to its scheduling policies and time-sensitiveness. The Cloud architecture is located in a different physical location than the industrial site where the actual control is done and the communication is performed through Internet. The results showed over 99% success rate for tasks requiring response within one second. They suggest that the architecture is feasible for soft or firm real-time applications.

Realizing an unproven concept in real environments without testing and validating requires good investment of engineering time and money. However, using virtual environments which can simulate several hours of real environment tasks in couple of minutes save a lot of time.

CloudSim is a framework to model and simulate Cloud Computing infrastructures and their services. It supports modelling and simulation of large scale Cloud data centers, their application containers, costs as well as power consumption [18]. One simulation tool to evaluate the reliability of the system is called iFogSim and implemented by Gupta, Dastjerdi, Ghosh, and Buyya [19]. It is based on CloudSim and allows addition of fog or edge devices, creation of topologies and evaluation of resource management policies focusing on latencies [19]. Sonmez, Ozgovde, and Ersoy introduced another simulator called EdgeCloudSim [20]. It adds a mobility model and non-fixed delays into the network which is fixed in iFogSim. The simulator also gives detailed information on resource usage as well as the percentage of tasks statuses.

In both simulators, the data is passed to the Cloud in case there are no resources available in the Edge/Fog Server. However, in our scenario, the Edge Servers can also offload the tasks to other Edge Servers by considering the available resources, network and computation delays. Additionally, the end-devices do not have mobility, only the data does. We believe that there are no available simulators in the literature which can offload the tasks of immobile end-devices between the Edge Servers nor a standard Edge Server architecture which is capable of performing real-time calculations. The aim in this research is not simply to build another architecture, but to analyse the existing architectures and consider industrial requirements to make up a generic reference architecture which is vendor-independent and extensible. This ongoing work will build a novel architecture comparing the existing architectures, initially simulating in a virtual platform. To the best of our knowledge, this is not considered in any of the aforementioned reference architectures.

III. CONCEPT

Although Cloud Computing reduces costs of computation by saving hardware and giving flexibility, the physical distance to the device reduces the QoS. Additionally, if the resources of a Cloud infrastructure is shared, scheduling the tasks is a difficult task. Moreover, transmitting too much data to the Cloud more than a network can handle is unnecessary and causes network congestion [21]. If the task execution is critical and time-constrained, then an in-time correct reaction is necessary.

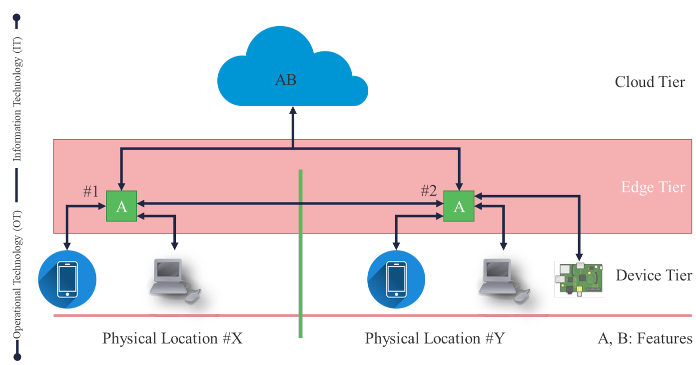


Figure 2. Simplified topology of the Edge Computing network.

One of the main goals of Edge Computing is to reduce latency and to keep the QoS as high as possible. As seen in Figure 1, in Cloud Computing, the Cloud infrastructure communicates with the end-devices directly. Edge Computing intends to solve the issues of Cloud Computing or IoT by adding an additional tier between the IoT devices and back-end infrastructure for computing and communication purposes. As depicted in Figure 2, this tier also has intermediate components for the first gathering, analysis, computation of the data. These intermediate components are called *Edge Servers*. Several architecture types for IoT-enabled applications are proposed

[22]. In this paper, a three-tier architecture is used. Unlike the example scenario of work done in [20], this paper assumes that the end-devices are not mobile.

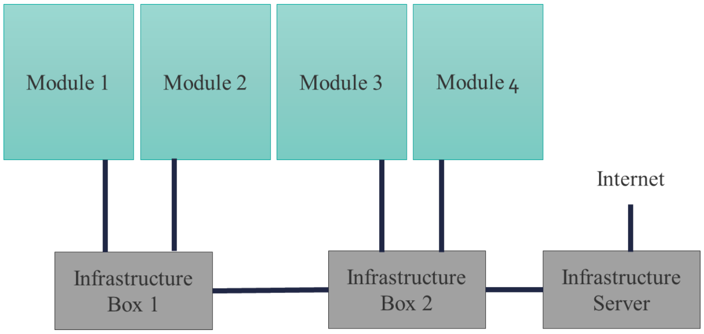


Figure 3. Overview of the modular testbed architecture to be used for validation and evaluation.

The scenario in our research involves a multi-vendor modular testbed for research purposes by *SmartFactory*^{KL} [23]. The testbed is composed of plug-and-produce modules and each of them performs one step of the production, independent from other modules. The modules are developed by different industrial partners and work together to produce a customizable and individualized product. As Figure 3 illustrates, the modules in the testbed are not directly communicating with each other, but through the infrastructure boxes. Each infrastructure box is connected with each other serially and provides pressured air, network connection, safety bus, and power to the modules. The communication to the Internet is performed through the central infrastructure server. The aim of the research is to add computing power into the infrastructure boxes to analyse, monitor the modules and react to the expected or unexpected situations, including real-time behaviour.

In our approach, we propose an extensible Edge Server model for Edge Computing to be integrated inside Infrastructure Boxes. If there are multiple Edge Servers in the same network, they are able to communicate with each other. Each server is orchestrated by itself, which means that they are not dependent of each other and aware of the neighbouring server capabilities in the same network. If a task cannot be guaranteed or performed by a server, the receiving server knows which other servers are capable of performing the same task. In all circumstances, the data transfer among devices will be performed through secured protocols. Figure 2 shows the simplified topology of a three-tier Edge Computing.

Edge Server is not a complete replacement of the Cloud with respect to its functionalities. Although its available resources are higher than the end-devices, they are lower than the Cloud. Instead, highly repeated tasks, or tasks that require in-time response are preferred to be executed in an Edge Server.

As seen in Figure 2, the proposed architecture for Edge Computing consists of *Cloud Tier*, *Edge Tier*, and *Device Tier*. In the Device Tier, there are end-user devices. The green blocks in the Edge Tier are Edge Servers. These servers gather, aggregate, analyse, and process the data before offloading them to the Cloud Tier or send back to the devices. The end-devices can be in the same physical location, or in different locations as depicted in the figure. When an end-device needs to communicate with the Cloud, first, the request is sent to the Edge Server which is at the closest location. Then, if the Edge Server is capable of completing the task by itself, it automatically handles the data and responds to the end-device with the result. If not, the data is offloaded to another server in the same tier provided that it exists. Otherwise, the data is offloaded to the Cloud. The decision process is made by considering available resources in other available servers in the same network, physical distance, and time requirements. In automation domain, Edge Tier can be seen as an edge or borderline between the Information Technology (IT) and Operational Technology (OT). In IT, the speed considerations are not critical whereas in OT, the communication or computing, or both must be real-time. Edge Tier isolates the network between IT and OT. Assume that A and B are features that could be serviced by the Cloud. For example, if a device in location X or Y needs the feature B to perform a task, the request will be orchestrated by the Edge Servers #1 or #2 and be sent to and performed by the Cloud. However, if, for example, the feature A is requested by an end-device in location X, first the Edge Server #1 will evaluate its own available resources. Depending on the urgency of the request, resource utilization, and calculated delays, it will either complete the request by itself or offload to the server #2 or to the Cloud.

Different tasks may have different priorities even though they are real-time. If there are multiple task requests, the server should pause the lower priority tasks while keeping track of the paused tasks or offloading them. The challenge here is to decide on the functionalities in the Edge Tier by keeping the costs at minimum and the QoS at maximum. However, deciding on the count and available resources of Edge Servers are also big challenges and big trade-offs. There are several aspects to consider before passing the data to the Cloud. To decide where to execute the task, each server has an orchestrator of which details will be explained in Section IV. According to this, the function should consider the priority of the task, resource utilization of the servers, computing cost for the task, and the physical distance or distance cost of the servers that is going to be used.

A. *Requirements*

Edge Computing is a paradigm which uses Cloud Computing technologies and gives more responsibilities to the Edge tier. These responsibilities are namely, computing offload, data caching/storage, data processing, service distribution, IoT management, security, and privacy protection [24].

Without limiting the Cloud Computing features, Edge Computing needs to have the following requirements, some of which are also defined for Cloud Computing [25][26]:

1) Interoperability: Servers in Edge Computing can connect with various devices and other servers. In Cloud Computing, IoT allows countless number of devices to communicate with humans or each other. This creates a big market for manufacturers of these devices. For this reason, there is the issue of interoperability with connected device using different communication protocols. Advanced Message Queuing Protocol (AMQP), Message Queue Telemetry Transport (MQTT), and TCP/IP are widely used and should be supported by Edge Computing. Using a widely-used and widely-known standard will remove the technology and language barriers, increasing interoperability among the devices.

2) Scalability: Similar to Cloud services, Edge Computing will also need to be adapted for the size of its users and sensors. Additional deployment of Edge Servers is costly and small number of Edge Servers is desirable in terms of economical aspects. For this reason, high scalability is also mandatory.

3) Extensibility: Computing technology is developing rapidly. After 2-3 years of deployment, clock speeds, memory size and program size increase, too. Easy deployment of new services and new devices with small effort is required for essential goal of Edge Computing. New functions and devices should be integrated without (re)configuration of the Edge network. Therefore, the system should allow extensibility with hardware and software components.

4) Abstraction: For the seamless control and communication, the abstraction of each Edge Node and group of nodes is required. Moreover, abstraction helps the topology of an Edge network to be flexible and reconfigurable. Fundamentally, an Edge node is located between device tier and Cloud tier. In other words, an Edge tier is a border between Information Technology (IT) and Operational Technology (OT). This tier can consist of one or more Edge nodes and groups. In this case, one Edge node of the group can share tasks or nodes in the group can be prioritized. Utilization of Application Programming Interfaces (APIs) in abstraction is useful to provide backward compatibility for the new functionalities or big changes in the architecture.

5) Time sensitiveness: Below OT, the operations may be near-real-time or real-time. Edge Computing is expected to solve time issues which Cloud computing cannot guarantee. Unlike Cloud Computing, physically close distance is one strength of reliable and fast communication without worrying about traffic problem. Video streaming service is one of expected applications of Edge Computing. It is required for real-timeness of the service provision. In addition, time-sensitiveness adds big benefits to the providers of reactive services, such as location-based advertisements and user-status based guide systems.

6) Security & Privacy: Using Cloud Computing services has a trade-off for enterprises like manufacturing and high-tech companies because there is a concern about the leakage of high knowledge and business activities outside their own organization. Edge Computing is a way to secure data contents, which is different from firewall which only controls external access into the network. It is also important to isolate the data by preventing access from even non-authorized users.

7) Reliability: Edge Servers provide real-time or non-real-time control for the devices. Real-time tasks may be vital which involve human safety. Therefore, it is vital to have a reliable system which reacts when it is needed and how it is needed. The physical reliability requirements for Edge servers providing services is similar to Cloud Computing. Harsh environments, such as factories and construction yards, require water-proof ceiling, fanless computers and dust-proof system. In power plant, magnetic shield is equipped by sensor gateways.

8) Intelligence: Multi-sensor generates tremendous amount of data and uploads into Cloud, directly. It causes network congestion and heavy load on the Cloud server. Edge Computing supports first and second filtering of these data by converting into higher level of data contents. Data filtering is implemented by rule-based engines or machine learning algorithms. In the case of multi-camera system like security systems, Edge Computing supports image processing, computer vision and enables object detection before transferring the data into the Cloud. Another example is predicting the failure or abnormalities in a production line by analysing the sensor data and taking the precautions for prevention or informing the user. These kinds of intelligent functions are necessary for Edge Computing.

9) Power: Unexpected shutdown or blackout is the cause of breakdown of Edge Server. Uninterruptible power supply (UPS) is required to give an ample amount of time to protect the electronic units and data storage in case of an unexpected shutdown due to power outage.

B. Enablers

Edge Computing uses wide range of technologies and brings them together. Within this domain, Edge Computing utilizes many technologies, such as wireless sensor networks (WSN), mobile data acquisition, mobile signature analysis, Fog/Grid Computing, distributed data operations, remote Cloud services, etc. Additionally, it combines the following protocols and terms:

1) 5G communication: It is the fifth generation wireless system which aims at higher capacity, lower power consumption, and lower latency compared to the previous generations. Due to increased amount of data between the data, 5G is expected to solve traffic issues which arose with the increased number of connected devices.

2) PLC protocols: Object Linking and Embedding for Process Control Unified Architecture (OPC-UA) is a protocol developed for industrial automation. Due to its openness and robustness, it is widely used by industries in the area of oil and gas, pharmaceutical, robotics, and manufacturing.

3) Message queue broker: MQTT and TCP/IP are popular message protocols of smart sensors and IoT devices. Supporting these message brokers, Edge Computing increases the device count that it connects. For the problem of MQTT security, AMQP is useful in the communication with Cloud Computing server.

4) Event processor: After messages of IoT arrive in the Edge server, event processor analyses those messages and creates semantic events using pre-defined rules. EsperNet, Apache Spark, and Flink are some examples for this enabler.

5) Virtualisation: Cloud services are deployed as virtual machines on a Cloud server or clusters. Using virtual machines allow running multiple instances of operating systems (OS) on the same server.

6) Hypervisor: As well as virtual machine, performance evaluation and data handling are required and realized by hypervisor to control virtual machines in the host computer.

7) OpenStack: Managing multiple resources could be challenging. OpenStack is a Cloud operating system that helps control of pools of computing and storage resources at ease through a control panel and monitoring tools.

8) AI platform: Rule-based engine and Machine learning platform supports data analysis in local level. As stated in Section III-A, this is quite important to reach one of the goals of Edge Computing which is to gather, analyse, and perform the first filtering of the data.

9) Docker: Virtual machines work with installation of operating systems. Unlike virtual machines, Docker is a Container as a Service (CaaS), which can use a single shared operating system and run software in isolated environment. It only requires the libraries of the software which makes it a lightweight system without worrying about where the software is deployed.

IV. Architecture Design

An Edge Server must be capable of gathering, aggregating the data, computing and transferring it back to the end-device. However, in the meantime, the servers must be able to communicate with other Edge Servers within the network, in case their resources are not enough to perform the task. Alternatively, they must be able to offload the data to the Cloud. In other words, the Edge Servers must have a reliable and communicable network between each other and the Cloud.

For seamless task handling and communication, the servers must follow some standards compatible with each other. This is not a simple task, since this technology contains several aspects to consider such as resource allocation, scheduling, scaling, storage, etc. The architecture must also be able to handle time-critical or real-time tasks. The software must also be designed or modified to work with the real-time capable system [27]. Last but not least, the server must be extensible with plug-and-play modules to advance or add new functionalities via hardware or software modules. These extensions must be validated before usage, to keep the system functional and to prevent intrusion. Such architecture design is divided into Hardware Modules and Software Components which are explained in the next subsections.

A. Hardware Modules

As mentioned in the previous sections, to solve the problems of Cloud Computing, Edge Servers or "Edge Nodes" need to have more computing power than the end-devices. However, the hardware in the Edge Tier is also limited compared to the Cloud. Therefore, to keep the balance between the performance and costs, first, the use cases for the Edge Servers must be defined, then, the resource must be considered to handle these defined use cases in-time. During the research, it has been decided to use Samsung Artik 710 as the hardware. It has high performance 8-core 64-bit ARM Processor, integrated wireless adapters, 1 GB RAM and 4 GB flash memory, extensible with an SD card. The price to performance ratio, internal real-time clock (RTC), and availability of the open source repository also played a big role for this decision [28][29].

The data produced by the end-devices are not directly sent to the Cloud or back-end infrastructure, but initial computing is performed on these servers. Considering the number of connected devices and the data they produced, these servers are used to aggregate, analyse, and process the data before sending it into the upper layer, the infrastructure, or back to the device.

The proposed Edge Server architecture is to be designed modular and should provide functionalities for real-time and non-real-time control, as well as real-time communication. Each server in our proposal has a *Core Node* of which functionalities can be extended by plugging additional optional modules in. Figure 4 shows an overview of possible modules or devices that could be attached to the core node. Hardware modules can also have their computing power, or simply improve the available functionality of the core node. For example, to capture big data and store, a storage module can be attached or in the case that machine learning algorithms are desired to be executed on the server, a dedicated artificial intelligence (AI) module with dedicated Graphics Processing Unit (GPU) can be connected. This module will be available for use with none to minimal configuration.

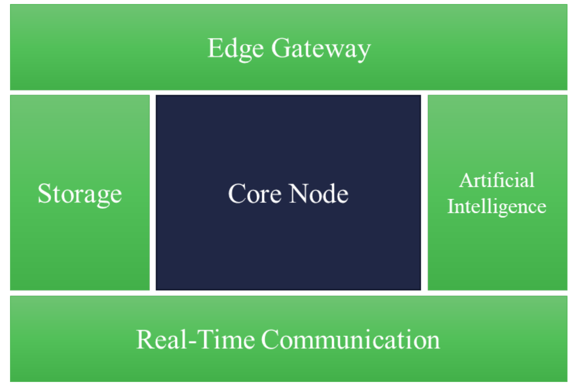


Figure 4. View of the proposed extensible Edge server architecture with its major functionalities, where green blocks extend the functionalities for the blue core node.

To preserve the integrity of the system and prevent unauthorized hardware from breaking the system functionality, only verified hardware must be allowed for connection. Hardware extensions are planned to be made using a physical master key, which is thought to be a USB device. Whenever this device is plugged in, the Edge Server will be ready to identify and allow new hardware to be added.

As mentioned in Section III-A, scalability is quite important to accomplish the tasks. In the scope of scalability, one server is expected to be aware of its neighbouring servers along with their functionalities. When a server is plugged into the network and turned on, first, it publishes that it is available. Then, it publishes its resources along with the available functionalities inside. Using the previous example, in case an AI module is connected to one server, other servers are informed with this functionality and they can utilize this server more often for AI-related tasks. The decision, of course, depends on the conditions required by the task, such as deadline.

Core Node should support multiple communication interfaces. Therefore, time and speed considerations are quite important for choosing the best hardware. The Edge Server is also expected to perform real-time computing and control. Therefore, reliability and stability of the hardware is also mandatory.

B. Software Components

As mentioned in Section IV-A, the server has a *Core Node* which is capable of performing real-time tasks. Before implementing the software architecture of such system, it is important to define the roles of software, clarify, and assign separate roles to the components. Rather than choosing the technologies or languages, which the software components are going to be implemented in, all components defined here can be implemented in any language as long as they satisfy the

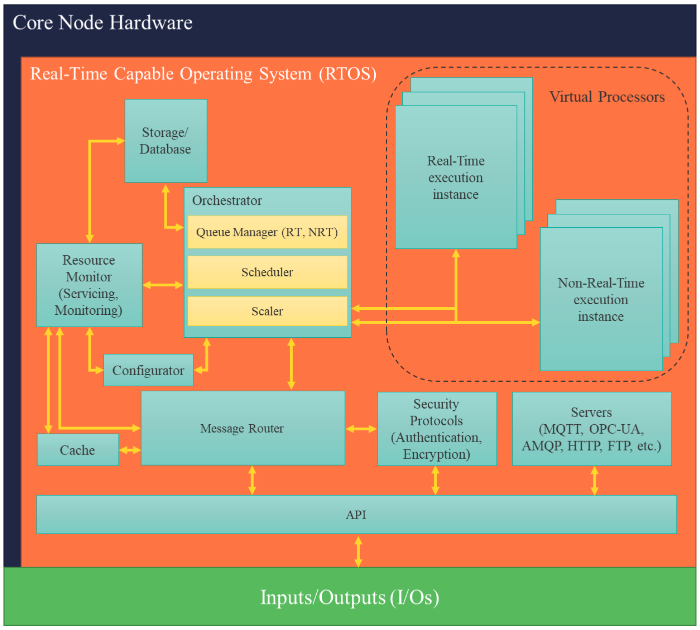


Figure 5. Proposed Software Architecture of Core Node.

requirements. Time-sensitiveness requires implementation of several components which are compatible with each other. These components should enable a reliable, stable in-time response and take correct actions.

Figure 5 shows the required software components inside the *Core Node* to fulfil the requirements. Below, these components are explained:

Real-Time Capable Operating System (RTOS): To implement software components, the operating system (OS) in the hardware must also be real-time capable. There are many OSes which can handle real-time tasks. However, the choice of the OS should be made by considering its available support, availability of the source code or openness for modifications, and applicability of the OS into the chosen hardware platform. Having a real-time capable kernel does not mean that the system will work in real-time. Applications, APIs, and the system must be designed properly to benefit from real-time functionality [30]. It should have a native support for the chosen *Core Node* hardware that it is running on. With the current stage of the research, the operating system has been decided to be based on Linux kernel, as it is open source and modification is easy with the plenty of resources available. The chosen Samsung Artik 710 hardware comes with pre-installed Fedora 24. However, the out-of-the-box version is not real-time capable. Therefore, the real-time patch [31] will be integrated into the kernel and it will be recompiled. Nevertheless, the other software components will be OS-neutral, therefore could be adapted to other OSes.

Inputs/Output (I/Os)s: I/Os are the interfaces which connect the hardware with the software. These are also used to connect other physical modules with the core node such as Edge Gateway for real-time communication.

API: APIs will be used for all communication with the I/Os, the end-devices, or the Cloud. An API is necessary to abstract the functionalities of other components. It allows internal modifications in case a new software component added without requiring complete change in the system. It also guarantees that the requests cannot interfere with the internal components since direct access to the individual modules or components is not allowed. Another goal of the architecture is to keep the migration efforts at minimum. Therefore, it is important to choose an accepted and standard language for the API to make it compatible with as many device and software as possible. Another aspect to consider is to choose a lightweight, yet stable API as a low latency is desired. Last but not least, the API should make sure that the requests are always authorized. This is performed by evaluating the request with *Security Protocols* component.

Message Router: As soon as the task or data arrives, this component retrieves and routes it to the location where task should be handled by communicating with *Resource Monitor* component. In case there are no resources available in this server, the task will either be transferred to another Edge Server or to the Cloud. This component always makes sure that the incoming task is from a trusted device by interacting with the *Security Protocols* component.

Configurator: The server and their modules are automatically configured as soon as they are attached. Nevertheless, their manual configuration or tweaks are performed via this component. It provides a Web-based and shell-based administrator panel to modify server properties, monitor the status, perform low-level resource allocations, and adjust orchestrator parameters, etc. Additionally, this component detects other nearby servers in the same network and configures the server to use them, when necessary. Similar to other components, this component is also accessed through the API.

Storage/Database: This component is used to store temporary data for the active or waiting tasks. However, the component will not keep the permanent data of the tasks. To keep the permanent data, storage module or the Cloud is recommended.

Servers: Standalone servers which are used out-of-the-box with only configuration changes are encapsulated in this component. The servers enable API communication as well as internal communication among the components. The servers are configured through the API via *Configurator* component.

Security Protocols: One of the most important requirement for the Edge Computing is keeping the data secure and private. Security itself is a vast aspect to consider. Therefore, a dedicated component to handle all security-related issues is necessary. This component monitors all incoming connections and takes the necessary actions in case the request is unexpected or unauthorized. Since the Edge Server will be accessible via the Cloud, the component is also responsible to prevent Internet-based attacks. Moreover, the component makes sure that the data privacy is preserved by encrypting, decrypting the data and issuing and exchanging keys, etc.

Resource Monitor (RM): The computing power in the Edge Server hardware is limited. RM actively monitors all available resources of Edge Servers in the network. It is also aware of all the other connected Edge Servers and their attached modules. RM directly interacts with the *Message Router* and decides whether the task received must be processed in this server or offloaded to another location. If the task is to be executed in another server, this component informs the

Message Router and points the target without further action. This decision is made by bi-directional communication with *Orchestrator*.

Cache: A temporary storage component to serve the data faster for the future requests. It is especially used when *Resource Monitor* decides that the task is not going to be executed in the current server.

Virtual Processors: For performance and power reasons, it is typical to have multi-core hardware. In multi-core hardware, one thread is generally executed on a single core. If a software is not optimized for multi-core, it cannot benefit from multi-core hardware. On the other hand, multi-threaded software execution is distributed among the cores. However, in either case, it is possible to set central processing unit (CPU) affinity of a process, that is a running instance of the software or program. Low-level programming makes configuration of the kernel possible to add virtual processors, limit or specify the available resources, and assign specific processes to these virtual processors.

Orchestrator: If RM allows execution of the task in this server, this component becomes active. Using event-based communication with the RM, the task is handled according to its urgency or the priority. If there is enough resource to execute the task, the task is immediately executed. If not, determined by the availability of the resource, task can be handled in different ways using the sub-components. This component has three different sub-components, namely: *Scheduler*, *Scaler*, and *Queue Manager*.

Scheduler: If a task is chosen to be executed in this server, it must be carefully scheduled to avoid deadline misses, especially for real-time tasks. Although multi-threaded software is usually scheduled by the OS and assigned to the CPU cores, the affinities may need to be adjusted. Depending on urgency or the priority of the new task, this component is responsible to set CPU affinities for the running tasks taking their priorities into consideration. The scheduler should be designed to minimize the waiting time of paused tasks.

Scaler: Scheduler sorts the execution times of the tasks. In the proposed architecture, some of the cores are dedicated for real-time tasks. Tasks not optimized for multi-core systems are by default assigned to run on a single core. If *Scheduler* is not able to meet the deadlines of the critical tasks, this component can increase the available core count for the real-time tasks to have them run on multiple cores, even if the cores are not assigned to execute real-time tasks. Of course, this is only possible provided that the tasks are multi-threaded.

Queue Manager: If a task cannot be executed immediately, one other possibility is to queue it. The queue contains both real-time (RT) and non-real-time (NRT) tasks. This component communicates with the *Scheduler*, and stores the tasks marked to pause and forwards the paused tasks to scheduler, following a variable scheduling algorithm.

V. Experiment Results on Different Hardware

SmartFactory[KL] concept was depicted in Figure 3 in Section III. It was also mentioned that, the factory comprises multiple modules from different vendors. Figure 6a shows the the modules, infrastructure boxes and the infrastructure server as they are deployed in real world. Figure 6b, shows one of the modules that is responsible for one task during production.

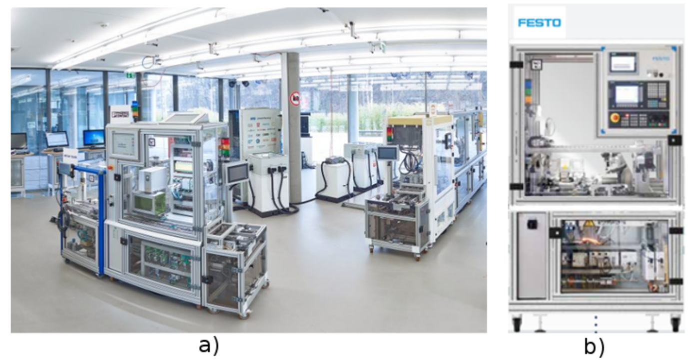


Figure 6. a) *SmartFactory*[KL] with all modules, infrastructure boxes and the infrastructure server, b) Close-up view of one of the modules.

Edge Server will be used in several real world scenarios such as object detection, production priority change, emergency stop and production cancellation. As a first experiment, we chose object detection use case in which we can analyze the objects using a smart glass. The experiment does not have a real-time dependency, therefore they were made with best effort approach. The experiment exposes an on-site Edge Server to improve the overall performance and realize new features which are not applicable in Cloud-computing.

It compares Edge Computing-based service with the only low-end device-based computing. Comparison with various combination of hardware is the experiment test coming from the problem of Cloud-based computing services. The test shows the advantage of using Edge Computing and help find the conditions where Edge Computing is better than Cloud or low-end device-based computing. The purpose of this experiment is to prove the improvement that Edge Computing configuration shows better performance than legacy configuration or Cloud-based system configuration. The experiment contains two computing servers which are CPS modules located in front of a machine itself and Edge Server in order to share the computing load. The test was made using five different network configurations. The configurations are divided into 1) CPS module alone, 2) Edge Server alone, 3) Connection with Edge server via wired communication, 4) Connection via wired and wireless communication, and 5) Connection via two different wireless communication. The input is video streaming data transferring the objects and people in front of the camera, attached in the production modules.

As seen in the Table I, we found the following results through the tests. Using CPS module with an Edge Server performed four times faster object detection than the CPS module alone. All kinds of connection types satisfied the bandwidth requirements of single connection with CPS module. Even if the wireless connection was 400 Mbps, the image streaming data could be transferred without any delay. Faster wireless connection will be more useful when there are more data being transferred in order to keep low latency. CPS test did not show any delay but the computation time was long since no Cloud services are used and computing power is weak. Edge Server alone showed the best results, however needed more memory because of the need to run additional software to test the performance.

Test Configuration	CPS Module	Comm. Method	Edge Computing
CPS module alone without Edge service (Raspberry Pi 3: Quad core 1.2GHz 64bit CPU and 1 GB RAM)	U: 4.6 to 4.83s C: 22 - 67% M: <360 KB (36%)	Not used	Not used
Edge Server alone (Intel I7 64bit 16GB RAM NVIDIA GTX 950)	Not used	Not used	U: 1.2 to 1.40s C: 28 - 51% M: <550∼650 MB G: <6∼16% (210-290MB)
CPS module communicating with Edge Server via Wired connection	Only transferring camera images	1 Gbps Wired Connection	U: 1-2s C: 16-23% M: 250-310 MB G: 6-17% (571MB)
CPS module communicating via 802.11ngb with Edge Server with the router via Wired connection	Only transferring camera images	802.11ngb @0.4 Gbps and 1 Gbps wired connection	U: 1-2s C: 17-23% M: 250-310 MB G: 6-14% (570MB)
CPS module communicating via 802.11ngb with Edge Server with the router via Wireless connection	Only transferring camera images	802.11ngb @0.4Gbps and 802.11ac @1.2Gbps	U: 1-2s C: 17-23% M: 250-310 MB G: 6-14% (570MB)

TABLE I. Object detection experiment results with different configurations (U: Update period (Average (AVG)), C: CPU Util. AVG, M: Memory Usage, G: Graphical Memory Load).

VI. Conclusion and Future Work

Edge Computing is a recent term which moves the services from the Cloud to the device as close as possible and open for new innovations. It is a borderline between the Cloud and the device tier. Although the Cloud Computing and IoT have brought many advantages in the previous years, increased number in the connected devices raised some issues, such as latency and low QoS problems. Edge Computing is believed to solve these issues by analysing the issues and considering the requirements of real world use cases.

There are already several existing proposed architectures in the domain of Edge Computing, such as EdgeX Foundry, Liota, and OpenFog Reference Architecture. Although they are also extensible and they allow inter-connectivity, they do not focus on the real-timeliness of the architectures.

This paper showed an ongoing work on how *Edge Computing* physically looks like together with its requirements and enablers. It also explained the basics on how the communication between the end-devices and Edge Servers are expected to be. Last, the paper proposed an ongoing work for Edge Server architecture which is capable of performing real-time computations. The servers are able to orchestrate the tasks and find the best host to offload the requested task by an end-device. The offloading can be done between other Edge Servers in the network, or the Cloud. The work is being developed by considering the real-world use cases of the industrial partners. The first experiment is made using object detection algorithms with intermediate devices, without real-time requirement. However, further experiments will be performed with industrial use cases requiring real-time deadlines. Later on, the comparison with the legacy systems will be made.

One ongoing task is implementation of a simulator similar to CloudSim framework, including all components explained in Section IV. This will help us find out the optimal parameters for the hardware modules and the software components. Current version of the simulator is capable of emulating the computer central processing unit (CPU), generating tasks with specific properties, and scheduling them with feasibility checks. Later, the architecture will be implemented, validated and evaluated on an optimal hardware chosen for the work.

In the proposed solution, when a task execution started, it can be performed only using one server. Another future work is to divide these tasks into multiple machines to exploit the resource utilization of the available resources, which is called task migration. This work could further be extended by using artificial intelligence instead of mathematical calculation for optimizations. For example, similar tasks can be grouped, and the historical data can be used to estimate the task duration and the action can be taken.

Acknowledgment

This research was funded in part by the H2020 program of European Union, project number 723094 (project FAR-EDGE). The responsibility for this publication lies with the authors. The project details can be found under project website at: `http://www.far-edge.eu`

References

[1] V. Gezer, J. Um, and M. Ruskowski, "An extensible edge computing architecture: Definition, requirements and enablers," in Eleventh International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies. International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM-2017), Special Session on Edge Computing, International Academy, Research, and Industry Association (IARIA). IARIA, November 2017. [Online]. Available: https://www.researchgate.net/publication/321134141_An_Extensible_Edge_Computing_Architecture_Definition_Requirements_and_Enablers

[2] NCTA, "The Growth of The Internet of Things," Infographic, May 2014, [retrieved: May 2018]. [Online]. Available: https://www.ncta.com/platform/industry-news/infographic-the-growth-of-the-internet-of-things/

[3] D. Evans, "The Internet of Things - Cisco," Cisco, White Paper, April 2011, [retrieved: May 2018]. [Online]. Available: https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

[4] H. H. Holm, J. M. Hjelmervik, and V. Gezer, "CloudFlow - an infrastructure for engineering workflows in the cloud," in UBICOMM 2016: The Tenth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies. IARIA, October 2016, pp. 158–165.

[5] P. M. Mell and T. Grance, "The nist definition of cloud computing," in National Institute of Standards and Technology Technical report, September 2011, [retrieved: May 2018]. [Online]. Available: https://www.nist.gov/publications/nist-definition-cloud-computing?pub_id=909616

[6] R. Lhotka, "Should all apps be n-tier?" Blog, 2005, [retrieved: May 2018]. [Online]. Available: http://www.lhotka.net/weblog/ShouldAllAppsBeNtier.aspx

[7] E. Olderog and H. Dierks, Real-Time Systems: Formal Specification and Automatic Verification. Cambridge University Press, 2008.

[8] T. Kaldewey, C. Lin, and S. Brandt, "Firm real-time processing in an integrated real-time system," Universitry of York, Department of Computer Science - Report, vol. 398, 2006, p. 5.

[9] T. Goldschmidt, M. K. Murugaiah, and C. Sonntag, "Cloud-Based Control: A Multi-Tenant, Horizontally Scalable Soft-PLC," in IEEE 8th International Conference on Cloud Computing, 2015.

[10] S. Higginbotham, "Sensor Networks Top Social Networks for Big Data," Article, 2010, [retrieved: May 2018]. [Online]. Available: https://gigaom.com/2010/09/13/sensor-networks-top-social-networks-for-big-data-2/

[11] T. Valich, “Big Data In Planes: New P&W Gtf Engine Telemetry To Generate 10GB/s,” Article, 2015, [retrieved: May 2018]. [Online]. Available: https://vrworld.com/2015/05/08/big-data-in-planes-new-pw-gtf-engine-telemetry-to-generate-10gbs/

[12] I. C. A. Center, “IBM: Internet of Things,” Cloud Garage Method, 2017, [retrieved: Sep 2017]. [Online]. Available: https://www.ibm.com/devops/method/content/architecture/iotArchitecture/industrie_40

[13] “OpenFog Consortium Reference Architecture,” Website, 2017, [retrieved: May 2018]. [Online]. Available: https://www.openfogconsortium.org/ra/

[14] “EdgeX Foundry Architectural Tenets,” EdgeX Foundry Wiki, 2017, [retrieved: May 2018]. [Online]. Available: https://wiki.edgexfoundry.org/display/FA/Introduction+to+EdgeX+Foundry

[15] “VMware Introduces Liota,” Website, 2017, [retrieved: May 2018]. [Online]. Available: https://www.vmware.com/radius/vmware-introduces-liota-iot-developers-dream/

[16] “pICASSO Project,” Website (German), [retrieved: May 2018]. [Online]. Available: https://www.projekt-picasso.de/projekt/

[17] O. Givehchi, J. Imtiaz, H. Trsek, and J. Jasperneite, “Control-as-a-service from the cloud: A case study for using virtualized plcs,” in 2014 10th IEEE Workshop on Factory Communication Systems (WFCS 2014), May 2014, pp. 1–4.

[18] “The Internet of Things - Cisco,” CloudSim Website, [retrieved: May 2018]. [Online]. Available: http://www.cloudbus.org/cloudsim/

[19] G. H., A. V. Dastjerdi, S. K. Ghost, and R. Buyya, “iFogSim: A Toolkit for Modeling and Simulation of Resource Management Techniques in Internet of Things, Edge and Fog Computing Environments,” in Software Practive and Experience, June 2016.

[20] C. Sonmez, A. Ozgovde, and C. Ersoy, “Edgecloudsim: An environment for performance evaluation of edge computing systems,” in 2017 Second International Conference on Fog and Mobile Edge Computing (FMEC), May 2017, pp. 39–44.

[21] H. Al-Bahadili, Simulation in Computer Network Design and Modeling: Use and Analysis: Use and Analysis, ser. Premier reference source. Information Science Reference, 2012.

[22] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, “Internet of things: A survey on enabling technologies, protocols, and applications,” IEEE Communications Surveys Tutorials, vol. 17, no. 4, 2015, pp. 2347–2376.

[23] D. Zuehlke, “Smartfactory — towards a factory-of-things,” vol. 34, no. 1, 2010, pp. 129 – 138. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1367578810000143

[24] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge Computing: Vision and Challenges,” IEEE Internet of Things Journal, vol. 3, no. 5, October 2016, pp. 637–646.

[25] G. Orsini, D. Bade, and W. Lamersdorf, “Context-Aware Computation Offloading for Mobile Cloud Computing: Requirements Analysis, Survey and Design Guideline,” Procedia Computer Science, vol. 56(1), December 2015, pp. 10–17.

[26] J. Shamsi, M. A. Khojaye, and M. A. Qasmi, “Data-intensive cloud computing: Requirements, expectations, challenges, and solutions,” Journal of Grid Computing, vol. 11, no. 2, Jun 2013, pp. 281–310.

[27] S. Rostedt, “Intro to Real-Time Linux for Embedded Developers,” Linux Foundation Blog, 2013, [retrieved: May 2018]. [Online]. Available: https://www.linuxfoundation.org/blog/intro-to-real-time-linux-for-embedded-developers/

[28] “Samsung ARTIK IoT Platform - ARTIK 710 IoT module,” Samsung IoT Website, [retrieved: May 2018]. [Online]. Available: https://www.artik.io/modules/artik-710/

[29] “Samsung ARTIK Kernel on GitHub,” Github, [retrieved: May 2018]. [Online]. Available: https://github.com/SamsungARTIK/linux-artik/

[30] J. Huang, “RTMux: A Thin Multiplexer to Provide Hard Realtime Applications for Linux,” Embedded Linux Conference Europe, October 2014, [retrieved: May 2018]. [Online]. Available: https://events.linuxfoundation.org/sites/events/files/slides/rtmux_1.pdf

[31] L. K. Projects, “RTLinux Repository,” Website, 2007, [retrieved: May 2018]. [Online]. Available: https://www.kernel.org/pub/linux/kernel/projects/rt/

Deduction System for Decision Logic Based on Many-valued Logics

Yotaro Nakayama* ‡, Seiki Akama† and Tetsuya Murai‡
*Nihon Unisys, Ltd., 1-1-1, Toyosu, Koto-ku, Tokyo, 135-8560, Japan
Email: yotaro.nakayama@unisys.co.jp
†C-Republic, Inc., 1-20-1 Higashi-Yurigaoka, Asao-ku, Kawasaki-shi, 215-0012, Japan
Email: akama@jcom.home.ne.jp
‡Chitose Institute of Science and Technology, 758-65 Bibi, Chitose, Hokkaido, 066-865, Japan
Email: t-murai@photon.chitose.ac.jp

Abstract—**Rough set theory has been extensively used both as a mathematical foundation of granularity and vagueness in information systems and in a large number of applications. However, the decision logic for rough sets is based on classical bivalent logic; therefore, it would be desirable to develop decision logic for uncertain, ambiguous and inconsistent objects. In this study, a deduction system based on partial semantics is proposed for decision logic. We propose Belnap's four-valued semantics as the basis for three-valued and four-valued logics to extend the deduction of decision logic since the boundary region of rough sets is interpreted as both a non-deterministic and inconsistent state. We also introduce the consequence relations to serve as an intermediary between rough sets and many-valued semantics. Hence, consequence relations based on partial semantics for decision logic are defined, and axiomatization by Gentzen-type sequent calculi is obtained. Furthermore, we extend the sequent calculi with a weak implication to hold for a deduction theorem and also show a soundness and completeness theorem for the four-valued logic for decision logic.**

Keywords–rough set; decision logic; consequence relation; many-valued logic; sequent calculi.

I. INTRODUCTION

This research paper is an extended version of an earlier paper [1] presented at the IARIA Conference on SEMAPRO 2017. Pawlak introduced the theory of rough sets for handling rough (coarse) information [2]. Rough set theory is now used as a mathematical foundation of granularity and vagueness in information systems and is applied to a variety of problems. In applying rough set theory, decision logic was proposed for interpreting information extracted from data tables. However, decision logic adopts the classical two-valued logic semantics. It is known that classical logic is not adequate for reasoning with indefinite and inconsistent information. Moreover, the paradoxes of the material implications of classical logic are counterintuitive.

Rough set theory can handle the concept of approximation by the indiscernibility relation, which is a central concept in rough set theory. It is an equivalence relation, where all identical objects of sets are considered elementary. Rough set theory is concerned with the lower and upper approximations of object sets. These approximations divide sets into three regions, namely, the positive, negative, and boundary regions. Thus, Pawlak rough sets have often been studied in a three-valued logic framework because the third value is thought to correspond to the boundary region of rough sets [3][4].

On the contrary, in this paper, we propose that the interpretation of the boundary region is based on four-valued semantics rather than three-valued since the boundary region can be interpreted as both undefined and overdefined. For example, a knowledge base K of a Rough set can be seen as a theory *KB* whose underlying logic is L. *KB* is called *inconsistent* when it contains theorems of the form A and $\sim A$ (the negation of A). If *KB* is not inconsistent, it is called *consistent*. Our approach for a rough set proposes useful theory to handle such inconsistent information without system failure. In this study, non-deterministic features are considered the characteristic of partial semantics. Undetermined objects in the boundary region of rough sets have two interpretations of both undefinedness and inconsistency.

The formalization of both three-valued and four-valued logics is carried out using a consequence relation based on partial semantics. The basic logic for decision logic is assumed to be many-valued, in particular, three-valued or four-valued and some of its alternatives [5]. If such many-valued logics are used as a basic deduction system for decision logic, it can be enhanced to a more useful method for data analysis and information processing. The decision logic of rough set theory will be axiomatized using Gentzen sequent calculi and a four-valued semantic relation as basic theory. To introduce many-valued logic to decision logic, consequence relations based on partial interpretation are investigated, and the sequent calculi of many-valued logic based on them are constructed. Subsequently, many-valued logics with weak implication are considered for the deduction system of decision logic.

The deductive system of decision logic has been studied from the granule computing perspective, and in [6], an extension of decision logic was proposed for handling uncertain data tables by fuzzy and probabilistic methods. In [7], a natural deduction system based on classical logic was proposed for decision logic in granule computing. In [3], the sequent calculi of the Kleene and Łukasiewicz three-valued logics were proposed for rough set theory based on non-deterministic matrices for semantic interpretation. The Gentzen-type axiomatization of three-valued logics based on partial semantics for decision logic is proposed in [1]. The reasoning for rough sets is comprehensively studied in [8].

The paper is organized as follows. In Section II, we briefly review rough sets, the decision table, and decision logic. In Section III, Belnap's four-valued semantics is introduced as the basis of the semantics interpretation presented in the paper. In Section IV, we present a partial semantics model for rough sets and decision logic based on four-valued semantics, and some characteristics are presented. In Section V, an axiomatization using Gentzen sequent calculus is presented according to a consequence relation based on the previously discussed partial semantics. In Section VI, we discuss the extension of sequent calculi for many-valued logics with weak negation and implication to enable a deduction theorem. In Section V II, A soundness and completeness theorem is showed for a four-valued sequent calculus GC4. Finally, in Section VIII, a summary of the study and possible directions for future work are provided.

II. Rough Sets and Decision Logic

Rough set theory, proposed by Pawlak [2], provides a theoretical basis of sets based on approximation concepts. A rough set can be seen as an approximation of a set. It is denoted by a pair of sets called the lower and upper approximations of the set. Rough sets are used for imprecise data handling. For the upper and lower approximations, any subset X of U can be in any of three states according to the membership relation of the objects in U. If the positive and negative regions on a rough set are considered to correspond to the truth-value of a logical form, then the boundary region corresponds to ambiguity in deciding truth or falsity. Thus, it is natural to adopt a three-valued logic.

Rough set theory is outlined below. Let U be a non-empty finite set called a universe of objects. If R is an equivalence relation on U, then U/R denotes the family of all equivalence classes of R, and the pair (U, R) is called a Pawlak approximation space. A knowledge base K is defined as follows:

Definition 1. A knowledge base K is a pair $K = (U, R)$, where U is a universe of objects, and **R** is a set of equivalence relations on the objects in U.

Definition 2. Let $R \in$ **R** be an equivalence relation of the knowledge base $K = (U, R)$ and X any subset of U. Then, the lower and upper approximations of X for R are defined as follows:

$$\underline{R}X = \bigcup\{Y \in U/R \mid Y \subseteq X\} = \{x \in U \mid [x]_{\mathrm{R}} \subseteq X\},$$
$$\overline{R}X = \bigcup\{Y \in U/R \mid Y \cap X \neq 0\} = \{x \in U \mid [x]_{\mathrm{R}} \cap X \neq \emptyset\}.$$

Definition 3. If K = (U,R), R ∈ R, and X ⊆ U, then the R-positive, R-negative, and R-boundary regions of X with respect to R are defined respectively as follows:

$$POS_R(X) = \underline{R}X,$$
$$NEG_R(X) = U - \overline{R}X,$$
$$BN_R(X) = \overline{R}X - \underline{R}X.$$

Objects included in an R-boundary are interpreted as the truth-value gap or glut. The semantic interpretation for rough sets is defined later.

Here, we denote the language of rough sets.

A. Decision Tables

Decision tables can be seen as a special important class of knowledge representation systems and can be used for applications. Let $K = (U, A)$ be a knowledge representation system and $C, D \subset A$ be two subsets of attributes called condition and decision attributes, respectively.

A KR-system with a distinguished condition and decision attributes is called a decision table, denoted $T = (U, A, V, s)$ or in short DC, where U is a finite and nonempty set of objects, A is a finite and nonempty set of attributes, V is a nonempty set of values for $a \in A$, and s is an information function that assigns a value $U \times s_x : A \to V$ (for simplicity, the subscript x will be omitted), where $\forall x \in U$, and $\forall a \in C \cup D \subset A$.

Equivalence classes of the relations $IND(C)$ and $IND(D)$, a subset of A, are called condition and decision classes, respectively.

With every $x \in U$, we associate a function $dx : A \to V$, such that $d_x(a) = a(x)$ for every $a \in C \cup D$; the function d_x is called a decision rule (in T), and x is referred as a label of the decision rule d_x.

The decision rule d_x is *consistent* (in T) if for every $y \neq x$, $d_x|C = d_y|C$ implies $d_x|D = d_y|D$; otherwise the decision rule is *inconsistent*.

A decision table is *consistent* if all of its decision rules are consistent; otherwise the decision table is *inconsistent*. Consistency (inconsistency) sometimes may be interpreted as determinism (non-determinism).

U	a	b	c	d	e
1	1	0	2	2	0
2	0	1	1	1	2
3	2	0	0	2	2
4	1	0	2	2	0
5	1	0	2	0	1
6	2	2	0	1	1
7	2	1	1	1	2
8	0	1	1	0	1

TABLE I. Decision table

Proposition 1. *A decision table $T = (U, A, V, s)$ is consistent iff $C \Rightarrow D$, where C and D are condition and decision attributes.*

From Proposition 1, it follows that the practical method of checking the consistency of a decision table is by simply computing the degree of dependency between the condition and decision attributes. If the degree of dependency equals 1, then we conclude that the table is consistent; otherwise, it is inconsistent.

Consider Table I from Pawlak [2]. Assume that a, b, and c are condition attributes and d and e are decision attributes. In this table, for instance, decision rule 1 is inconsistent, whereas decision rule 3 is consistent. Decision rules 1 and 5 have the same condition, but their decisions are different.

B. Decision Logic

A decision logic language (DL-language) L is now introduced [2]. The set of attribute constants is defined as $a \in A$, and the set of attribute value constants is $V = \bigcup V_a$. The propositional variables are φ and ψ, and the propositional connectives are $\perp$, $\sim$, $\wedge$, $\vee$, $\to$ and $\equiv$.

Definition 4. The set of formulas of the decision logic language (DL-language) L is the smallest set satisfying the following conditions:

1) (a, v), or in short a_v, is an atomic formula of L.
2) If φ *and* ψ are formulas of the DL-language, then $\sim\varphi$, $\varphi \wedge \psi$, $\varphi \vee \psi$, $\varphi \to \psi$, and $\varphi \equiv \psi$ are formulas.

The interpretation of the DL-language L is performed using the universe U in $S = (U, A)$ of the Knowledge Representation System (*KR-system*) and the assignment function, mapping from U to objects of formulas. Formulas of the DL-language are interpreted as subsets of objects consisting of a value v and an attribute a.

Atomic formulas (a, v) describe objects that have a value v for the attribute a. Attribute a is a function from U to V, defined by $a(x) = s_x(a)$, where $x \in U$, and $s_x(a) \in V$. If let $s_x(a) = v$, then a can be viewed as a binary relation on U, such that for $\langle x, v\rangle \in U \times U, \langle a, v\rangle \in a$ if and only if $a(x) = v$. In this case, the atomic formula (a, v) can be denoted by $a(x, v)$, where x is a variable, and v is taken as a constant; they are all terms in U. Thus, (a, v) can be viewed as formula $a(x, v)$ which is an atomic formula. The semantics for *DL* is given by a model. For *DL*, the model is the KR-system $S = (U, A)$, which describes the meaning of symbols of predicates (a, v) in U, and if we properly interpret the formulas in the model, then each formula becomes a meaningful sentence, expressing the properties of some objects. An object $x \in U$ satisfies a formula φ in $S = (U, A)$, denoted $x \models_S \varphi$ or in short $x \models \varphi$, iff the following conditions are satisfied:

Definition 5. The semantic relations of a DL-language are defined as follows:

$x \models_S a(x, v)$ iff $a(x) = v$,

$x \models_S {\sim}\varphi$ iff $x \not\models_S \varphi$,

$x \models_S \varphi \vee \psi$ iff $x \models_S \varphi$ *or* $x \models_S \psi$,

$x \models_S \varphi \wedge \psi$ iff $x \models_S \varphi$ *and* $S \models_S \psi$,

$x \models_S \varphi \to \psi$ iff $x \models_S {\sim}\varphi \vee \psi$,

$x \models_S \varphi \equiv \psi$ iff $x \models_S \varphi \to \psi$ and $s \models_S \psi \to \varphi$.

If φ is a formula, then the set $|\varphi|_S$ defined as follows:

$$|\varphi|_S = \{x \in U \mid x \models_S \varphi\}$$

and will be called the meaning of the formula φ in S. The following properties are obvious:

Proposition 2. The meaning of an arbitrary formula satisfies the following:

$|\neg\varphi|_S = U - |\varphi|_S$,

$|\varphi \vee \psi|_S = |\varphi|_S \cup |\psi|_S$,

$|\varphi \wedge \psi|_S = |\varphi|_S \cap |\psi|_S$,

$|\varphi \to \psi|_S = (U - |\varphi|_S) \cup |\psi|_S$,

$|\varphi \equiv \psi|_S = |\varphi|_S \to |\psi|_S \cap |\varphi|_S \to |\psi|_S$.

Thus, the meaning of the formula φ is the set of all objects having the property expressed by the formula φ, or the meaning of the formula φ is the description in the *KR*-language of the set objects $|\varphi|$. A formula φ is said to be true in a KR-system S, denoted $\models_S \varphi$, iff $|\varphi|_S = U$, i.e., the formula is satisfied by all objects of the universe in the system S. Formulas φ and ψ are equivalent in S iff $|\varphi|_S = |\psi|_S$.

Proposition 3. The following are the simple properties of the meaning of a formula.

$\models_S \varphi$ *iff* $|\varphi|_S = U$,

$\models_S {\sim}\varphi$ *iff* $|\varphi|_S = \emptyset$,

$\varphi \to \psi$ *iff* $|\psi|_S \subseteq |\psi|_S$,

$\varphi \equiv \psi$ *iff* $|\psi|_S = |\psi|_S$.

To deal with deduction in *DL*, we need suitable axioms and inference rules. Here, the axioms will correspond closely to the axioms of classical propositional logic, but some specific axioms for the specific properties of knowledge representation systems are also needed. The only inference rule will be *modus ponens*. We will use the following abbreviations:

$$\varphi \wedge {\sim}\varphi =_{def} 0 \text{ and } \varphi \vee {\sim}\varphi =_{def} 1.$$

A formula of the form

$(a_1, v_1) \wedge (a_2, v_2) \wedge \ldots \wedge (a_n, v_n)$,

where $v_{ai} \in V_a, P = \{a_1, a_2, \ldots, a_n\}$, and $P \subseteq A$, is called a *P-basic formula* or in short *P*-formula. An atomic formula is called an *A-basic formula* or in short a basic formula.

Let $P \subseteq A$, φ be a *P*-formula, and $x \in U$. The set of all *A*-basic formulas satisfiable in the knowledge representation system $S = (U, A)$ is called the *basic knowledge* in S. We write $\underset{s}{\Sigma}(P)$ to denote the disjunction of all *P*-formulas satisfied in S. If $P = A$, then $\underset{s}{\Sigma}(A)$ is called the *characteristic formula* of S.

The knowledge representation system can be represented by a data table. Its columns are labeled by attributes, and its rows are labeled by objects. Thus, each row in the table is represented by a certain *A*-basic formula, and the whole table is represented by the set of all such formulas. In *DL*, instead of tables, we can use sentences to represent knowledge. There are specific axioms of *DL*:

1. $(a, v) \wedge (a, u) \equiv 0$ for any $a \in A, u, v \in V$, and $v \neq u$.
2. $\bigvee_{v \in V_a} \equiv 1$ for every $a \in A$.
3. $\sim (a, v) \equiv \bigvee_{a \in V_a, u \neq v} (a, u)$ for every $a \in A$.

We say that a formula φ is derivable from a set of formulas Ω, denoted $\Omega\varphi$, iff it is derivable from the axioms and formulas of Ω by a finite application of modus ponens. Formula φ is a theorem of *DL*, denoted φ, if it is derivable from the axioms only. A set of formulas Ω is consistent iff the formula $\varphi \wedge {\sim}\varphi$ is not derivable from Ω. Note that the set of theorems of *DL* is identical with the set of theorems of classical propositional logic with specific axioms (1)–(3), in which negation can be eliminated.

Formulas in the KR-language can be represented in a special form called a *normal form*, which is similar to that in classical propositional logic. Let $P \subseteq A$ be a subset of attributes and let φ be a formula in the KR-language. We say that φ is in a P-normal form in S, in short in P-normal form, iff either φ is 0 or φ is 1, or φ is a disjunction of non-empty P-basic formulas in S. (The formula φ is non-empty if $|\varphi| \neq \emptyset$).

A-normal form will be referred to as normal form. The following is an important property in the *DL*-language.

Proposition 4. Let φ be a formula in a *DL*-language, and let P contain all attributes occurring in φ. Moreover, assume axioms (1)–(3) and the formula $\underset{s}{\Sigma}(A)$. Then, there is a formula ψ in the P-normal form such that $\varphi \equiv \psi$.

Definition 6. A translation τ from the propositional constant L to an interpretation of a rough set language $\mathcal{L}_{\mathrm{RS}}$ of atomic expressions in the *KR*-system S is combined with $\neg, \vee, \wedge$ and $\rightarrow$ such that

$$\tau(|\varphi|_S) = |(a,v)|_S,$$
$$\tau(|\sim\varphi|_S) = -\tau(|\varphi|_S),$$
$$\tau(|\varphi \vee \psi|_S) = \tau(|\varphi|_S) \cup \tau(|\psi|_S),$$
$$\tau(|\varphi \wedge \psi|_S) = \tau(|\varphi|_S) \cap \tau(|\psi|_S),$$
$$\tau(|\varphi \rightarrow \psi|_S) = -\tau(|\varphi|_S) \cup \tau(|\psi|_S),$$
$$\tau(|\varphi \equiv \psi|_S) = (\tau(|\varphi|_S) \cap \tau(|\psi|_S)) \cup (-\tau(|\varphi|_S) \cap -\tau(|\psi|_S)).$$

Let φ be an atomic formula of the DL-language, $R \in C \cup D$ an equivalence relation, X any subset of U, and a valuation v of propositional variables. Then, the truth-values of φ is defined as follows:

$$\|\varphi\|^v = \begin{cases} \mathbf{t} \ \textit{if} \ |\varphi|_S \subseteq POS_R(U/X) \\ \mathbf{f} \ \textit{if} \ |\varphi|_S \subseteq NEG_R(U/X) \end{cases}$$

This shows that decision logic is based on bivalent logic. In the next section, an interpretation of decision logic based on three-valued logics will be discussed.

III. Belnap's Four-Valued Logic

Belnap [9] first claimed that an inference mechanism for a database should employ a certain four-valued logic. The important point in Belnap's system is that we should deal with both incomplete and inconsistent information in databases. To represent such information, we need a four-valued logic since classical logic is not appropriate for the task. Belnap's four-valued semantics can in fact be viewed as an intuitive description of the internal states of a computer.

In Belnap's four-valued logic **B4**, four kinds of truth-values are used from the set **4** = $\{\mathbf{T}, \mathbf{F}, \mathbf{N}, \mathbf{B}\}$. These truth-values can be interpreted in the context of a computer, namely **T** means just told True, **F** means just told False, **N** means told neither True nor False, and **B** means told both True and False. Intuitively, **N** can be equated as $\emptyset$, and **B** as overdefined.

Belnap outlined a semantics for **B4** using logical connectives. Belnap's semantics uses a notion of *set-ups* mapping atomic formulas into **4**. A set-up can then be extended for any formula in **B4** in the following way:

$$s(A \ \& \ B) = s(A) \ \& \ s(B),$$
$$s(A \ \vee \ B) = s(A) \ \vee \ s(B),$$
$$s(\sim A) = \ \sim s(A).$$

Belnap also defined a concept of entailments in **B4**. We say that A entails B just in case for each assignment of one of the four values to variables, the value of A does not exceed the value of B in B4, i.e., $s(A) \leq s(B)$ for each set-up s. Here, $\leq$ is defined as $\mathbf{F} \leq \mathbf{B}$, $\mathbf{F} \leq \mathbf{N}$, $\mathbf{B} \leq \mathbf{T}$, $\mathbf{N} \leq \mathbf{T}$. Belnap's four-valued logic in fact coincides with the system of *tautological entailments* due to Anderson and Belnap [10]. Belnap's logic **B4** is one of the paraconsistent logics capable of tolerating contradictions. Belnap also studied the implications and quantifiers in **B4** in connection with question-answering systems. However, we will not go into detail here.
The structure that consists of these four elements and the five basic operators is usually called **B4**.

Designated elements and models: The next step in using **B4** for reasoning is to choose its set of *designated* elements. The obvious choice is $\mathcal{D} = \{\mathbf{T}, \mathbf{B}\}$ since both values intuitively represent a formula known to be true. The set $\mathcal{D}$ has the property that $a \wedge b \in \mathcal{D}$ iff both a and b are in $\mathcal{D}$, while $a \vee b \in \mathcal{D}$ iff either a or b is in $\mathcal{D}$. From this point, various semantics notions are defined on **B4** as natural generalizations of similar classical notions.

IV. Rough Sets and Partial Semantics

Partial semantics for classical logic has been studied by van Benthem in the context of the *semantic tableaux* [11][12].

This insight can be generalized to study consequence relations in terms of a Gentzen-type sequent calculus. To handle an aspect of vagueness on the decision logic, the forcing relation for the partial interpretation is defined as a four-valued semantic.

As the proposed approach can replace the base bivalent logic of decision logic, alternative versions of decision logic based on many-valued logics are obtained.

The model $\mathcal{S}$ of decision logic based on four-valued semantics consists of a universe U for the language L and an assignment function s that provides an interpretation for L. For the domain $|\mathcal{S}|$ of the model $\mathcal{S}$, a subset is defined by $S = \langle S^+, S^- \rangle$. The first term of the ordered pair denotes the set of *n*-tuples of elements of the universe that *verify* the relation S, whereas the second term denotes the set of n-tuples that *falsify* the relation.

The interpretation of the propositional variables of L for the model $\mathcal{S}$ is given by $S_{\mathcal{S}} = \langle (S)^+_{\mathcal{S}}, (S)^-_{\mathcal{S}} \rangle$. An *interpretation function* for a domain $|\mathcal{S}|$ in the standard way as a function s with domain L such that $s(x) \in |\mathcal{S}|^n$ if S is a relation symbol. We need two interpretation functions for each model here; a model for partial logic for a predicate symbol is a triple $\langle |\mathcal{S}|, s^+, s^- \rangle$, where s^+ and s^- are interpretation functions for $|\mathcal{S}|$. The denotation of a relation symbol consists of those tuples for which it is *true* that they stand in the relation; the antidenotation consists of the tuples for which this is *false*. As before, truth and falsity are neither true nor false, or it may be both true and false that some tuple stands in a certain relation. The following definition is modified from [13].

Definition 7. Partial Relation: An *n*-ary *partial relation* S on the domain $|\mathcal{S}_1|, \ldots, |\mathcal{S}_n|$ is a tuple $\langle S^+, S^- \rangle$ of the relations $S^+, S^- \subseteq |\mathcal{S}_1| \times \ldots \times |\mathcal{S}_n|$. The relation S^+ is called S's *denotation*; S^- is called S's *antidenotation*, $|\mathcal{S}_1| \times \ldots \times |\mathcal{S}_n| / (S^+ \cup S^-)$ its gap, and $S^+ \cap S^-$ its glut. A partial relation is *coherent* if its glut is empty, *total* if its gap is empty, *incoherent* if it is not coherent and *classical* if it is both coherent and total. A unary partial relation is called a *partial set*.

Definition 8. Partial Operation for **4**: Let $S_1 = \langle S_1^+, S_1^- \rangle$ and $S_2 = \langle S_2^+, S_2^- \rangle$ be partial relations. Define

$$-S_1 := \langle S_1^+, S_1^- \rangle \ (partial\ complementation),$$
$$S_1 \cap S_2 := \langle S_1^+ \cap S_2^-, S_1^+ \cup S_2^- \rangle \ (partial\ intersection),$$
$$S_1 \cup S_2 := \langle S_1^+ \cup S_2^-, S_1^+ \cap S_2^- \rangle \ (partial\ union),$$
$$S_1 \subseteq S_2 := \langle S_1^+ \subseteq S_2^-, S_1^+ \subseteq S_2^- \rangle \ (partial\ inclusion).$$

Partial inclusion means S_1 approximates S_2. Let A be some set of partial relations; then, following properties hold:

$$\bigcap A := \langle \bigcap\{S^+ \mid S \in A\}, \bigcup\{S^- \mid S \in A\}\rangle,$$
$$\bigcup A := \langle \bigcup\{S^+ \mid S \in A\}, \bigcap\{S^- \mid S \in A\}\rangle.$$

To handle three-valued and four-valued logic in a unified manner, we adopt the four-value interpretation by Belnap [9].

Let $\mathbf{4} = \{\mathbf{T}, \mathbf{F}, \mathbf{N}, \mathbf{B}\}$ be the truth-values for the four-valued semantics of L, where each value is interpreted as true, false, neither true nor false, and both true and false.

A model $\mathcal{S}$ determines a four-valued assignment v on atomic formula in the following way:

$$\|\varphi\|^v = \left\{\begin{array}{c}\mathbf{T}\\ \mathbf{F}\\ \mathbf{N}\\ \mathbf{B}\end{array}\right\} \text{ if } |\varphi, \sim\varphi|_S \cap S = \left\{\begin{array}{c}\{\varphi\}\\ \{\sim\varphi\}\\ \{\emptyset\}\\ \{\varphi, \sim\varphi\}\end{array}\right\}.$$

Then, the truth-values of φ on $S = (U, A)$ is defined as follows:

$$\|\varphi\|^v = \begin{cases}\mathbf{T} \text{ if } |\varphi|_S \subseteq POS_R(U/X)\\ \mathbf{F} \text{ if } |\varphi|_S \subseteq NEG_R(U/X)\\ \mathbf{N} \text{ if } |\varphi|_S \not\subseteq POS_R(U/X) \cup NEG_R(U/X)\\ \mathbf{B} \text{ if } |\varphi|_S \subseteq BN_R(U/X)\end{cases}$$

Definition 9 (Partial Model). A partial model for a propositional DL-language L is a tuple $\mathcal{M} = (\mathcal{T}, \mathcal{D}, \mathcal{O})$, where

- $\mathcal{T}$ is a non-empty set of truth-values.
- $\emptyset \subset \mathcal{D} \subseteq \mathcal{T}$ is the set of designated values.
- For every *n-ary connective* $\diamond$ of L, $\mathcal{O}$ includes a corresponding n-ary function $\tilde{\diamond}$ from $\mathcal{T}^n$ to $\mathbf{4}$.

Let W be the set of *well-formed formulas* of L. A (legal) valuation in a Partial Model $\mathcal{S}$ is a function $V : W \to \mathbf{4}$ that satisfies the following condition:

$$V(\diamond(\psi_1, \cdots, \psi_n)) \in \tilde{\diamond}(V(\psi_1), \cdots, V(\psi_n))$$

for every n-ary connective $\diamond$ of L and any $\psi_1, \cdots, \psi_n \in W$.

Let $\mathcal{V}_M$ denote the set of all valuations in the partial model $\mathcal{D}$. The notions of satisfaction under a valuation, validity, and consequence relation are defined as follows:

- A formula $\varphi \in W$ is satisfied by a valuation $v \in \mathcal{V}_M$, in symbols, $\mathcal{M} \models_v \varphi$, $v(\varphi) \in \mathcal{D}$.
- A sequent $\Sigma = \Gamma \Rightarrow \Delta$ is satisfied by a valuation $v \in \mathcal{V}_M$, in symbols, $\mathcal{M} \models_v \Sigma$, iff either v does not satisfy some formula in Γ or v satisfies some formula in Δ.
- A sequent Σ is valid, in symbols, $\models \Sigma$, if it is satisfied by all valuations $V \in \mathcal{V}_M$.
- The consequence relation on W defined by $\mathcal{M}$ is the relation $\mathcal{M} \vdash$ on sets of formulas in W such that, for any $T, S \subseteq W, T \vdash_{\mathcal{M}} S$ iff there exist finite sets $\Gamma \subseteq T, \Delta \subseteq S$ such that the sequent $\Gamma \Rightarrow \Delta$ is valid.

Definition 10. (Tarski truth definition for partial propositional logic) Let L be a set of propositional constants and let $v : P \to \{\mathbf{T}, \mathbf{F}, \mathbf{N}, \mathbf{B}\}$ be a (*valuation*) *function*.

$$\|p\|^v = v(p) \text{ if } p \in P$$

The truth-values of φ on the information system $S = (U, A)$ are represented by forcing relations as follows:

$$\|\varphi\|^v = \mathbf{T} \textit{ iff } \mathcal{M} \models_v^+ \varphi \textit{ and } \mathcal{M} \not\models_v^- \varphi,$$
$$\|\varphi\|^v = \mathbf{F} \textit{ iff } \mathcal{M} \not\models_v^+ \varphi \textit{ and } \mathcal{M} \models_v^- \varphi,$$
$$\|\varphi\|^v = \mathbf{N} \textit{ iff } \mathcal{M} \not\models_v^+ \varphi \textit{ and } \mathcal{M} \not\models_v^- \varphi,$$
$$\|\varphi\|^v = \mathbf{B} \textit{ iff } \mathcal{M} \models_v^+ \varphi \textit{ and } \mathcal{M} \models_v^- \varphi.$$

A semantic relation for the model $\mathcal{M}$ is defined following [11][14][13]. The truth and falsehood of a formula of the DL-language are defined in a model $\mathcal{M}$. The truth (denoted by $\models_v^+$) and the falsehood (denoted by $\models_v^-$) of the formulas of the decision logic in $\mathcal{M}$ are defined inductively.

Definition 11. The semantic relations of $\mathcal{M} \models_v^+ \varphi$ and $\mathcal{M} \models_v^- \varphi$ are defined as follows:

$\mathcal{M} \models_v^+ \varphi$ *iff* $\varphi \in M^+$,
$\mathcal{M} \models_v^- \varphi$ *iff* $\varphi \in M^-$,
$\mathcal{M} \models_v^+ \sim \varphi$ *iff* $\mathcal{M} \models_v^- \varphi$,
$\mathcal{M} \models_v^- \sim \varphi$ *iff* $\mathcal{M} \models_v^+ \varphi$,
$\mathcal{M} \models_v^+ \varphi \vee \psi$ *iff* $\mathcal{M} \models_v^+ \varphi$ *or* $\mathcal{M} \models_v^+ \psi$,
$\mathcal{M} \models_v^- \varphi \vee \psi$ *iff* $\mathcal{M} \models_v^- \varphi$ *and* $\mathcal{M} \models_v^- \psi$,
$\mathcal{M} \models_v^+ \varphi \wedge \psi$ *iff* $\mathcal{M} \models_v^+ \varphi$ *and* $\mathcal{M} \models_v^+ \psi$,
$\mathcal{M} \models_v^- \varphi \wedge \psi$ *iff* $\mathcal{M} \models_v^- \varphi$ *or* $\mathcal{M} \models_v^- \psi$,
$\mathcal{M} \models_v^+ \varphi \to \psi$ *iff* $\mathcal{M} \models_v^- \varphi$ *or* $\mathcal{M} \models_v^+ \psi$,
$\mathcal{M} \models_v^- \varphi \to \psi$ *iff* $\mathcal{M} \models_v^+ \varphi$ *and* $\mathcal{M} \models_v^- \psi$.

The symbol $\sim$ denotes strong negation, in which $\sim$ is interpreted as true if the proposition is false.

Since validity in **B4** is defined in terms of truth preservation, the set of designated values is $\{\mathbf{T}, \mathbf{B}\}$ of $\mathbf{4}$. We assume that an interpretation of **B4** satisfies the following constraint.

Definition 12. Exclusion and Exhaustion:

Exclusion: model $\mathcal{M}$ is *exclusion iff* $S^+ \cap S^- = \emptyset$.
Exhaustion: model $\mathcal{M}$ is *exhaustion iff* $S^+ \cup S^- = S$.

The model $\mathcal{M}$ is *consistent* if and only if $S^+ \cap S^- = \emptyset$. The relational domains of general models are closed under the operations $\cap, \cup$.

The natural operation on the set of truth combinations $\mathbf{4} = \{\mathbf{T}, \mathbf{F}, \mathbf{N}, \mathbf{B}\}$ that we have defined in the previous section can be extended to the class of partial relations.

Definition 13. A model of **B4** for L is a pair $M = (S, |\cdot|)$, where S is a non-empty set, and $|\cdot|$ is an interpretation of a propositional symbol, with $|p| : S_n \to \mathbf{4}$ for any $p \in P_n, n \leq 0$.

Example 1. Suppose the decision table below where the condition and decision attributes are not considered.

$U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$
Attribute: $C = \{c1, c2, c3, c4\}$
$c_1 = \{x_1, x_4, x_8\}, c_2 = \{x_2, x_5, x_7\}, c_3 = \{x_3\}$,
$c_4 = \{x_6\}$
$U/C = c_1 \cup c_2 \cup c_3 \cup c_4$
Any subset $X = \{x_3, x_6, x_8\}$

$POS_C(X) = c_3 \cup c_4 = \{x_3, x_6\}$
$BN_C(X) = c_1 = \{x_1, x_4, x_8\}$
$NEG_C(X) = c_2 = \{x_2, x_5, x_7\}$

The evaluation of the truth-values of the formulas is as follows:

If $|C_{c3}|_S \subseteq POS_C(X)$ *then* $\|C_{c3}\|^v = \mathbf{T}$,
If $|C_{c2}|_S \subseteq NEG_C(X)$ *then* $\|C_{c2}\|^v = \mathbf{F}$,
If $|C_{c2}|_S \not\subseteq POS_C(X) \cup NEG_C(X)$ *then* $\|C_{c2}\|^v = \mathbf{N}$,
If $|C_{c1}|_S \subseteq BN_C(X)$ *then* $\|C_{c1}\|^v = \mathbf{B}$.

Example 2. Consider Table I again. Assume that a, b, and c are condition attributes and d and e are decision attributes. Decision rules 1 and 5 are inconsistent. This means that 1 and 5 can be considered to have non-deterministic value, e.g., **N** or **B** respectively.

V. Consequence Relation and Sequent Calculus

Partial semantics in classical logic is closely related to the interpretation of the Beth tableau [12]. Van Benthem [11] suggested the relationship of the consequence relation to a Gentzen sequent calculus. We replace the bivalent logic of the decision logic with many-valued logics based on partial semantics.

A. Sequent Calculi for Many-valued Logics

We begin by recalling the basic idea of the Beth tableau. The Beth tableau proves $X \to Y$ by constructing a counterexample of $X \ \& \sim Y$. The Beth tableaux has several partial features. For instance, there may be counterexamples even if a branch remains open. This insight led van Benthem [11] to work out partial semantics for classical logic.

Here, we describe a brief introduction of sequent calculi. For sequent calculi, formulas are constructed from the propositional variables and logical connectives, e.g., $\sim$, $\neg$, $\wedge$, $\vee$, *and* $\to$. Capital letters $A, B, \ldots$ are used for formulas, and Greek capital letters Γ, Δ are used for finite sequences of formulas. A sequent is an expression of the form $\Gamma \Rightarrow A$. We introduce some concepts of sequent calculi. If a sequent $\Gamma \Rightarrow A$ is provable in a system S, then we write $S \vdash \Gamma \Rightarrow A$. A rule R of inference holds for a system S if the following condition is satisfied. For any instance of the following sequent of R, if $S \vdash \Gamma_i \Rightarrow A_i$ for all i, then $S \vdash \Delta \Rightarrow B$.

$$\frac{\Gamma_1 \Rightarrow A_1 \ \ldots \ \Gamma_n \Rightarrow A_n}{\Delta \Rightarrow B}$$

Moreover, R is said to be derivable in S if there is a derivation from $\Gamma_1 \Rightarrow A_1, \ldots, \Gamma_n \Rightarrow A_n$ to $\Delta \Rightarrow B$ in S.

To accommodate the Gentzen system to partial logics, we need some concepts of partial semantics. In the Beth tableau, It is assumed that V is a partial valuation function assigning the values 0 or 1 to an atomic formula p. We can then set $V(p) = 1$ for p on the left-hand side and $V(p) = 0$ for p on the right-hand side in an open branch of the tableau. To deal with an uncertain concept in many-valued semantics, we need to introduce the consequence relation [5]. *Pre* and *Cons* represent the sequent premise and conclusion, respectively, and 1 represents true and 0 false. First, we define the following concept of consequence relation C1.

(C1) for all V, if $V(Pre) = 1$, then $V(Cons) = 1$.

In C1, if *Pre* is evaluated as 1, then *Cons* preserves 1. Here, we define a classical Gentzen system.

Definition 14. The sequent calculus for the classical propositional logic CL is defined as follows:

Axiom: $A \Rightarrow A \ \ (ID)$

Sequent rules:

$$\frac{\Gamma \Rightarrow \Delta}{A, \Gamma \Rightarrow \Delta, A} \ (Weakening)$$

$$\frac{\Gamma, A \Rightarrow \Delta \quad \Gamma \Rightarrow A, \Delta}{\Gamma \Rightarrow \Delta} \ (Cut)$$

$$\frac{A, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, \sim A} \ (\sim R) \qquad \frac{\Gamma \Rightarrow \Delta, A}{\sim A, \Gamma \Rightarrow \Delta} \ (\sim L)$$

$$\frac{\Gamma \Rightarrow \Delta, A \quad \Gamma \Rightarrow \Delta, B}{\Gamma \Rightarrow \Delta, A \wedge B} \ (\wedge R) \qquad \frac{A, B, \Gamma \Rightarrow \Delta}{A \wedge B, \Gamma \Rightarrow \Delta} \ (\wedge L)$$

$$\frac{\Gamma \Rightarrow \Delta, A, B}{\Gamma \Rightarrow \Delta, A \vee B} \ (\vee R) \qquad \frac{A, \Gamma \Rightarrow \Delta \quad B, \Gamma \Rightarrow \Delta}{A \vee B, \Gamma \Rightarrow \Delta} \ (\vee L)$$

$$\frac{\Gamma \Rightarrow \Delta, \sim A, B}{\Gamma \Rightarrow \Delta, A \to B} \ (\to R)$$

$$\frac{\Gamma, \sim A \Rightarrow \Delta \quad B, \Gamma \Rightarrow \Delta}{A \to B, \Gamma \Rightarrow \Delta} \ (\to L)$$

Theorem 5. The logic for C1 is axiomatized by the Gentzen sequent calculus CL.

Proof: See [12],[11],[15]. ■

Next, we define the sequent calculus GC1 for C1 that can be obtained by adding the following rules to CL without $(\sim R)$ such as CL$\backslash\{(\sim R)\}$, where, "$\backslash$" implies that the rule following "$\backslash$" is excluded:

$$\frac{\Gamma \Rightarrow \Delta, A}{\Gamma \Rightarrow \Delta, \sim\sim A} \ (\sim\sim R) \qquad \frac{A, \Gamma \Rightarrow \Delta}{\sim\sim A, \Gamma \Rightarrow \Delta} \ (\sim\sim L)$$

$$\frac{\Gamma \Rightarrow \Delta, \sim A, \sim B}{\Gamma \Rightarrow \Delta, \sim(A \wedge B)} \ (\sim\wedge R)$$

$$\frac{\sim A, \Gamma \Rightarrow \Delta \quad \sim B, \Gamma \Rightarrow \Delta}{\sim (A \wedge B), \Gamma \Rightarrow \Delta} \ (\sim\wedge L)$$

$$\frac{\Gamma \Rightarrow \Delta, \sim A \quad \Gamma \Rightarrow \Delta, \sim B}{\Gamma \Rightarrow \Delta, \sim (A \vee B)} \ (\sim \vee R)$$

$$\frac{\sim A, \sim B, \Gamma \Rightarrow \Delta}{\sim(A \vee B), \Gamma \Rightarrow \Delta} \ (\sim \vee L)$$

It is worth noting that the three-valued logic by Kleene has no tautology. Thus, to define a consequence relation, a tableau system for a three-valued logic is formalized [11] [15]. Then, the consequence relation C2 is defined as follows:

(C2) for all V, if $V(Pre) = 1$, then $V(Cons) \neq 0$.

C2 is interpreted as *exclusion*; then, the consequence relation C2 is regarded for Kleene's strong three-valued logic K_3. As the semantics for C2, we define the extension of the valuation function $V^{C2}(p)$ for an atomic formula p as follows:

$\mathbf{T} =_{def} V^{C2}(p) = 1 =_{def} V^{C2}(p) = 1 \ and \ V^{C2}(p) \neq 0$,
$\mathbf{F} =_{def} V^{C2}(p) = 0 =_{def} V^{C2}(p) = 0 \ and \ V^{C2}(p) \neq 1$,

$\mathbf{N} =_{def} V^{C2}(p) = \{\} =_{def} V^{C2}(p) \neq 1 \ and \ V^{C2}(p) \neq 0.$

The interpretation of C2 by the partial semantics is given as follows:

Definition 15. $\Gamma \models_s \varphi$ iff there is no φ that is not **F** under V^{C2} (in the three-valued $\{\mathbf{T},\mathbf{F},\mathbf{N}\}$) and for all $\gamma \in \Gamma$, γ is **T** under V^{C2}.

The Gentzen-type sequent calculus GC2 axiomatizes C2 [15][11]. We are now in a position to define GC2. For GC2, the principle of explosion (ex falso quodlibet (EFQ)), defined below, is added to TG1$\backslash\{(\sim L)\}$.

(EFQ) $A, \sim A \Rightarrow$

Definition 16. The sequent calculus GC2 is defined as follows:
GC2 := $\{(ID), (Weakening), (Cut), (EFQ), (\wedge R), (\wedge L), (\vee R), (\vee L), (\rightarrow R), (\rightarrow L), (\sim\sim R), (\sim\sim L), (\sim\wedge R), (\sim\wedge L), (\sim\vee R), (\sim\vee L)\}$.

GC2 can be interpreted as *truth preserving* with the matrix of a three-valued logic defined as $\langle\{T, F, N\}, \{T\}, \{\sim, \vee, \wedge, \rightarrow\}\rangle$. For the rule $(\sim L)$ obtained from (EFQ), GC2 and GC1 are equivalent.

Theorem 6. GC2 = GC1.

Proof: (EFQ) can be considered as $(\sim L)$; then, double negation and the de Morgan laws in GC2 are obtained. ■

In the classical interpretation of CL, the law of excluded middle (EM) holds but not in C2.

Then, the rule C2 for the Gentzen system is axiomatized as GC2.

Theorem 7. C2 can be axiomatized by the sequent calculus GC2.

Proof: See [11][15]. ■

Theorem 8. In the model for C2, $\mathcal{S}$, DL-language L, and formula φ, it is not the case that $\mathcal{M} \models_v^+ \varphi$ and $\mathcal{M} \models_v^- \varphi$ hold.

Proof: Only the proof for $\sim$ and $\wedge$ will be provided. It can be carried out by induction on the complexity of the formula. The condition of *consistent* implies that it is not the case that $\varphi \in \mathcal{S}^+$ and $\varphi \in \mathcal{S}^-$. Then, it is not the case that $\mathcal{M} \models_v^+ \varphi$ and $\mathcal{M} \models_v^- \varphi$.

$\sim$: We assume that $\mathcal{M} \models_v^+ \sim \varphi$ and $\mathcal{M} \models_v^- \sim \varphi$ hold. Then, it follows that $\mathcal{M} \models_v^+ \varphi$ and $\mathcal{M} \models_v^- \varphi$. This is a contradiction.

$\wedge$: We assume that $\mathcal{M} \models_v^- \varphi \wedge \psi$ and $\mathcal{M} \models_v^+ \varphi \wedge \psi$ hold. Then, it follows that $\mathcal{M} \models_v^+ \varphi$, $\mathcal{M} \models_v^+ \psi$ and either $\mathcal{M} \models_v^- \varphi$ or $\mathcal{M} \models_v^- \psi$. In either case, there is a contradiction. ■

Next, we provide another consequence relation with a different interpretation for the third-value below.

(C3) for all V, if $V(Pre) \neq 0$, then $V(Cons) = 0$.

C3 is interpreted as *exhaustion*, then the consequence relation C3 is for Logic for Paradox [16]. As the semantics for C3, we define the extension of the valuation function $V^{C3}(p)$ for an atomic formula p as follows:

$\mathbf{T} =_{def} V^{C3}(p) = 1 =_{def} V^{C3}(p) = 1 \ and \ V^{C3}(p) \neq 0,$
$\mathbf{F} =_{def} V^{C3}(p) = 0 =_{def} V^{C3}(p) = 0 \ and \ V^{C3}(p) \neq 1,$
$\mathbf{B} =_{def} V^{C3}(p) = \{1,0\} =_{def} V^{C3}(p) = 1 \ and \ V^{C3}(p) = 0.$

The interpretation of C3 by the partial semantics is given as follows:

Definition 17. $\Gamma \models_v \varphi$ iff there is φ that is **T** under V^{C3} (in the three-valued $\{\mathbf{T},\mathbf{F},\mathbf{B}\}$) and for all $\gamma \in \Gamma$, γ is not **F** under V^{C3}.

The Gentzen sequent calculus GC3 is obtained from GC2, replacing EFQ with EM (*excluded middle*) as an axiom:

(EM) $\Rightarrow A, \sim A$

Definition 18. The sequent calculus GC3 is defined as follows:
GC3 := $\{(ID), (Weakening), (Cut), (EM), (\wedge R), (\wedge L), (\vee R), (\vee L), (\rightarrow R), (\rightarrow L), (\sim\sim R), (\sim\sim L), (\sim\wedge R), (\sim\wedge L), (\sim\vee R), (\sim\vee L)\}$.

Theorem 9. C3 can be axiomatized by the Gentzen calculus GC3.

Proof: GC3 can be obtained by deriving double negation and two de Morgan laws in GC3. The $(\sim R)$ rule can be provided as EM. ■

Next, we extend consequence relation C4 as follows:

(C4) *for all* V, *if* $V(Pre) \neq 0$, *then* $V(Cons) \neq 0$.

C4 is regarded as a four-valued logic since it allows for an inconsistent valuation. We are now in a position to define Belnap's four-valued logic **B4**.

As the semantics for GC4, Belnap's **B4** is adopted here. We define the extension of the valuation function $V^{C4}(p)$ for an atomic formula p as follows:

$\mathbf{T} =_{def} V^{C4}(p) = 1 =_{def} V^{C4}(p) = 1 \ and \ V^{C4}(p) \neq 0,$
$\mathbf{F} =_{def} V^{C4}(p) = 0 =_{def} V^{C4}(p) = 0 \ and \ V^{C4}(p) \neq 1,$
$\mathbf{N} =_{def} V^{C4}(p) = \{\} =_{def} V^{C4}(p) \neq 1 \ and \ V^{C4}(p) \neq 0,$
$\mathbf{B} =_{def} V^{C4}(p) = \{1,0\} =_{def} V^{C4}(p) = 1 \ and \ V^{C4}(p) = 0.$

The interpretation of C4 by the partial semantics is given as follows:

Definition 19. $\Gamma \models_v \varphi$ *iff* there is no φ that is not **F** under V^{C4} (in **4**) and for all $\gamma \in \Gamma$, γ is not **F** under V^{C4}.

Definition 20. The sequent calculus GC4 is defined as follows:
GC4 := $\{(ID), (Weakening), (Cut), (\wedge R), (\wedge L), (\vee R), (\vee L), (\sim\sim R), (\sim\sim L), (\sim\wedge R), (\sim\wedge L), (\sim\vee R), (\sim\vee L)\}$.

Theorem 10. C4 can be axiomatized by the sequent calculus GC4.

Proof: GC4 can be obtained by deriving double negation and two de Morgan laws in GC4. The $(F\sim)$ rule can be provided as EM. ■

VI. Extension of Many-valued Semantics

We introduce three-valued logics and provide some relationship and properties between the consequence relations we denotated in the previous section.

Kleene's strong three-valued logic: Kleene proposed three-valued logics to deal with *undecidable* sentences in connection with recursive function theory [17]. Thus, the third truth-value can be interpreted as undecided in the strong Kleene logic K_3, which is of special interest to describe a machine's computational state. K_3 can give a truth value to a compound sentence even if some of its parts have no truth value. Kleene also proposed the weak three-valued logic in which the whole sentence is undecided if any component of a compound sentence is undecided.

The truth tables for K_3 are defined as follows:

$\sim$	T	F	N
	F	T	N

$\wedge$	T	F	N
T	T	F	N
F	F	F	F
N	N	F	N

$\vee$	T	F	N
T	T	T	T
F	T	F	N
N	T	N	N

$\rightarrow$	T	F	N
T	T	F	N
F	T	T	T
N	T	N	N

The implication $\rightarrow$ can be defined in the following way:

$$A \rightarrow B =_{def} \sim A \vee B$$

The axiomatization of K_3 by a Gentzen-type sequent calculus can be found in the literature [5].

Let $\models$ be the consequence relation of K_3. Then, we have the following Gentzen-type sequent calculus GK_3 for K_3, which contains an axiom of the form

$X \models Y \text{if} X \cap Y \neq \emptyset$

and the rules (Weakening), (Cut), and

$$A \models \sim\sim A, \quad \sim\sim A \models A, \quad A, \sim A \models,$$
$$A, B \models A \wedge B, \quad A \wedge B \models A, \quad A \wedge B \models B,$$
$$\sim A \models \sim(A \wedge B), \quad \sim B \models \sim(A \wedge B),$$
$$\sim(A \wedge B) \models \sim A, \sim B.$$

GC2 is considered as Kleene's strong three-valued logic K_3. The implication of K_3 does not satisfy the deduction theorem. In addition, $A \rightarrow A$ is not a theorem in K_3.

Theorem 11. $\models_{K3} = \vdash_{C2}$, where $\models_{K3}$ denotes the consequence relation of K_3.

Proof: By induction on K_3 and GC2. It is easy to transform each proof of K_3 into GC2. The converse transformation can be also presented. ■

Łukasiewicz three-valued logic: Łukasiewicz's (1920) three-valued logic was proposed in order to interpret a future contingent statement in which the third truth-value can be read as indeterminate or possible. Thus, in Łukasiewicz's three-valued logic L_3, neither the law of excluded middle nor the law of non-contradiction holds. The difference between K_3 and L_3 lies in the interpretation of implication, as the truth table indicates.

It is also possible to describe the Hilbert presentation of L_3. Let $\supset$ be the Łukasiewicz implication. Then, we can show the following axiomatization of L_3 due to Wajsberg. It has been axiomatized by Wajsberg (1993) in [5] using a language based on $(\vee, \supset, \sim)$, the modus ponens rule and the following axioms:

(W1) $(p \supset q) \supset ((p \supset r) \supset (p \supset r))$,
(W2) $(\sim p \supset \sim q) \supset (q \supset p))$,
(W3) $(((p \supset \sim p) \supset p) \supset p)$.

They are closed under the rules of substitution and modus ponens. Unlike in K_3, $A \supset A$ is a theorem in L_3. It is noted, however, that the philosophical motivation of L_3 in connection with Aristotelian logic can be challenged. For a review of various three-valued logics, see Urquhart [5].

$\supset$	T	F	N
T	T	F	N
F	T	T	T
N	T	N	T

The definition of the semantic relation for the implication of L_3 is obtained by replacing the implication in Definition 11 with the following definition:

$\mathcal{M} \models_v^+ \varphi \rightarrow \psi$ *iff* $\mathcal{M} \models_v^- \varphi$ *or* $\mathcal{M} \models_v^+ \psi$ or
$(\mathcal{M} \not\models_v^+ \varphi$ *and* $\mathcal{M} \not\models_v^- \varphi$ *and* $\mathcal{M} \not\models_v^+ \psi$ *and* $\mathcal{M} \not\models_v^- \psi)$.
$\mathcal{M} \models_v^- \varphi \rightarrow \psi$ *iff* $\mathcal{M} \models_v^+ \varphi$ *and* $\mathcal{M} \models_v^- \psi$.

Logic of Paradox (LP): Logic of Paradox (LP) has been studied by Priest [16], which is one of the paraconsistent logics excluding EFQ. As motivation for paraconsistent logics in general, LP can treat various logical paradoxes and Dialetheism, which is a philosophical position that admits some true contradictions. GC3 is taken as a sequent calculus of LP [18], and the truth table of LP can be obtained K_3's truth value **N** replaced with **B**.

The definition of the semantic relation for the implication of GC3 is obtained by replacing the implication in Definition 11 with the following definition.

$\mathcal{M} \models_v^+ \varphi \rightarrow \psi$ *iff* $\mathcal{M} \not\models_v^+ \varphi$ *or* $\mathcal{M} \not\models_v^- \psi$ or
$(\mathcal{M} \models_v^+ \varphi$ *and* $\mathcal{M} \models_v^- \varphi$ *and* $\mathcal{M} \models_v^+ \psi$ *and* $\mathcal{M} \models_v^- \psi)$.
$\mathcal{M} \models_v^- \varphi \rightarrow \psi$ *iff* $\mathcal{M} \models_v^+ \varphi$ *and* $\mathcal{M} \models_v^- \psi$.

Belnap's four-valued logic: In section III, we have already seen the Belnap's four-valued logic. In addition to section III, we define the truth tables for $\sim$, $\wedge$, and $\vee$.

In this paper, the implication of B4 is defined with $\sim$ and $\vee$ and it does not hold for the rule of *modus ponens* because the *disjunctive syllogism* does not hold.

$\sim$	T	F	N	B
	F	T	N	B

$\wedge$	T	F	N	B
T	T	F	N	B
F	T	T	T	T
N	T	T	T	T
B	T	F	N	B

$\vee$	T	F	N	B
T	T	F	N	B
F	T	T	T	T
N	T	T	T	T
B	T	F	N	B

The aim of this paper is to present many-valued semantics for the decision logic. There are three candidates of consequence relations for the enhancement in the decision logic. GC2, which was discussed above, is interpreted as strong Kleene three-valued logic. The value of a proposition is neither true nor false in GC2. In this case, the designated value of GC2 is defined as $\{\mathbf{T}, \mathbf{N}\}$. GC3 is a paraconsistent logic, and its designated valued is defined as $\{\mathbf{T}, \mathbf{B}\}$. The paraconsistent logic does not hold for the principle of explosion (ex falso quodlibet); therefore, it is possible to interpret the consequence relation by C3. GC4 is obtained from C4 based on four-valued semantics and interpreted as both paracomplete and paraconsistent.

Here, we present the extended version of many-valued logics with weak negation $\neg$. Weak negation represents the lack of truth. The assignment of weak negation is defined as follows:

$$\|\neg\varphi\|_s = \begin{cases} \mathbf{T} \ \mathit{if} \ \|\varphi\|_s \neq \mathbf{T} \\ \mathbf{F} \ \mathit{otherwise} \end{cases}$$

Weak implication is defined as follows:

$$A \rightarrow_w B =_{def} \neg A \vee B$$

The assignment of weak implication is defined as follows:

$$\|A \rightarrow_w B\|^s = \begin{cases} \|B\|^s \ \mathit{if} \ \|A\|^s \in \mathcal{D} \\ \mathbf{T} \ \mathit{if} \ \|A\|^s \notin \mathcal{D} \end{cases}$$

We represent the truth tables for $\neg$ and $\rightarrow_w$ below.

$\neg$	T	F	N	B
	F	T	T	F

$\rightarrow_w$	T	F	N	B
T	T	F	N	B
F	T	T	T	T
N	T	T	T	T
B	T	F	N	B

The semantic relation for weak negation is as follows:

$$\begin{array}{lll} \mathcal{M} \models_v^+ \neg\varphi & \text{iff} & \mathcal{M} \not\models_v^+ \varphi, \\ \mathcal{M} \models_v^- \neg\varphi & \text{iff} & \mathcal{M} \models_v^+ \varphi. \end{array}$$

We try to extend many-valued logics with weak negation and weak implication. This regains some properties that some many-valued logics lack, such as the rule of modus ponens and the decision theorem. Obviously, L_3 recovers some properties that K_3 lacks and L_3's implication and weak implication has a close relationship.

Weak negation can represent the absence of truth. However, $\sim$ can serve as strong negation to express the verification of falsity. Note also that weak implication obeys the deduction theorem. This means that it can be regarded a logical implication. We can also interpret weak negation in terms of strong negation and weak implication:

$$\neg A =_{def} A \rightarrow_w {\sim}A$$

We define the sequent rules for $(\neg)$ and $(\rightarrow_w)$ as follows:

$$\frac{\Gamma, A \Rightarrow \Delta}{\Gamma \Rightarrow \neg A, \Delta}\ (\neg R) \qquad \frac{\Gamma \Rightarrow A, \Delta}{\Gamma, \neg A \Rightarrow \Delta}\ (\neg L)$$

$$\frac{A, \Gamma \Rightarrow \Delta, B}{\Gamma \Rightarrow \Delta, A \rightarrow_w B}\ (\rightarrow_w R) \qquad \frac{B, \Gamma \Rightarrow \Delta \quad \Gamma \Rightarrow \Delta, A}{A \rightarrow_w B, \Gamma \Rightarrow \Delta}\ (\rightarrow_w L)$$

GC2, GC3, and GC4 have additional rules of weak negation and weak implication, and we obtain $\mathrm{GC2}^+$, $\mathrm{GC3}^+$, and $\mathrm{GC4}^+$. $\mathrm{GC2}^+$ is the same as the extended Kleene logic EKL, that was proposed by Doherty [19] as the underlying three-valued logic for the non-monotonic logic and is provided with the deduction theorem.

$\mathrm{GC4}^+$ is interpreted as both paracomplete and paraconsistent. This prevents the paradox of material implication of classical logic.

Here, it is observed that L_3 can be naturally interpreted in $\mathrm{GC2}^+$. The Łukasiewicz implication can be defined as

$$A \supset B =_{def} (A \rightarrow_w B) \wedge ({\sim}B \rightarrow_w {\sim}A)$$

Next, we present the interpretation of weak negation for consequence relations C2, C3, and C4, which are interpreted as K_3, LP, and B4, respectively.

$$\|\neg A\|^{C2} = \begin{cases} \mathbf{F} \text{ if } \|A\| = \mathbf{T} \\ \mathbf{T} \text{ if } \mathit{otherwise} \end{cases}$$

$$\|\neg A\|^{C3} = \begin{cases} \mathbf{T} \text{ if } \|A\| = \mathbf{F} \\ \mathbf{F} \text{ if } \mathit{otherwise} \end{cases}$$

$$\|\neg A\|^{C4} = \begin{cases} \mathbf{F} \text{ if } \|A\| = \mathbf{T} \ \mathit{or} \ \mathbf{B} \\ \mathbf{T} \text{ if } \|A\| = \mathbf{F} \ \mathit{or} \ \mathbf{N} \end{cases}$$

We consider an application of weak negation for the interpretation of rough sets.

Example 3. Suppose the definition of a decision table is the same as Example 1.

$U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$

Attribute: $C = \{c1, c2, c3, c4\}$
$c_1 = \{x_1, x_4, x_8\}, c_2 = \{x_2, x_5, x_7\}, c_3 = \{x_3\}$,
$c_4 = \{x_6\}$
$U/C = c_1 \cup c_2 \cup c_3 \cup c_4$

Any subset $X = \{x_3, x_6, x_8\}$

$POS_C(X) = c_3 \cup c_4 = \{x_3, x_6\}$

$BN_C(X) = c_1 = \{x_1, x_4, x_8\}$

$NEG_C(X) = c_2 = \{x_2, x_5, x_7\}$

The interpretation of the consequence relation C4 for weak negation in the decision table is defined as follows:

If $|C_{c3}|_S \subseteq POS_C(X)$ *then* $\neg\|C_{c3}\|^v = \mathbf{F}$,

If $|C_{c2}|_S \subseteq NEG_C(X)$ *then* $\neg\|C_{c2}\|^v = \mathbf{T}$,

If $|C_{c2}|_S \not\subseteq POS_C(X) \cup NEG_C(X)$ *then* $\neg\|C_{c2}\|^v = \mathbf{T}$,

If $|C_{c1}|_S \subseteq BN_C(X)$ *then* $\neg\|C_{c1}\|^v = \mathbf{F}$.

VII. Soundness and Completeness

The soundness and completeness theorem is shown for the sequent system GC4$^+$. Other systems can be adopted in a similar way for GC4$^+$. GC4$^+$, which was discussed above, is interpreted as one Belnap's four-valued logic B4 extended with weak negation and weak implication. The sequent calculus GC4$^+$ is defined as follows:

$$\begin{aligned}\text{GC4}^+ := \{&(ID),\ (Weakening),\ (Cut),\ (\wedge R),\ (\wedge L),\\ &(\vee R),\ (\vee L),\ (\sim\sim R),\ (\sim\sim L), (\sim\wedge R),\ (\sim\wedge L),\\ &(\sim\vee R),\ (\sim\vee L),\ (\neg R),\ (\neg L),\ (\rightarrow_w R),\ (\rightarrow_w L)\}\end{aligned}$$

It is assumed that GC4$^+$ is the basic deduction system for decision logic obtained from C4 with weak negation and weak implication. This prevents the paradox of material implication of classical logic.

As the semantics for GC4$^+$, Belnap's **B4** is adopted here; we obey the definition of the valuation function $V^{C4}(p)$.

Lemma 12. The validity of the inference rules

1) The axioms of GC4$^+$ are valid.
2) For any inference rules of GC4$^+$ and any valuation s, if s satisfies all of the formulas of *Pre*, then s satisfies *Cons*.

Proof: 1) In GC4$^+$, the axiom (ID) and structural rules (weakening) and (cut) preserve validity.

For 2), the proof for $(\neg R)$, $(\rightarrow_w L)$, and $(\sim\wedge L)$ will be provided.

$(\neg R)$:

$$\frac{\Gamma, A \Rightarrow \Delta}{\Gamma \Rightarrow \neg A, \Delta}\ (\neg R)$$

Suppose that $\models_{GC4^+} \Gamma, A \Rightarrow \Delta$. Then, either (1) $v(\gamma) \neq \mathbf{T}$ for some $\gamma \in \Gamma$ or $v(\delta) \neq \mathbf{F}$ for some $\delta \in \Delta$ or (2) $v(A) \neq \mathbf{T}$. If (1) holds, then clearly $\models_{GC4^+} \Gamma \Rightarrow \neg A, \Delta$ *iff* $\models^-_{GC4^+} \Gamma$ *or* $\models^+_{GC4^+} \Delta, \neg A$. If (2) holds, then from the definition of $\neg$, it follows that $v(\neg A) = \mathbf{T}$ and then $\models_{GC4^+} \Gamma \Rightarrow \neg A, \Delta$.

$(\rightarrow_w L)$:

$$\frac{B, \Gamma \Rightarrow \Delta \quad \Gamma \Rightarrow \Delta, A}{A \rightarrow_w B, \Gamma \Rightarrow \Delta}\ (\rightarrow_w L)$$

Suppose that $\models_{GC4^+} B, \Gamma \Rightarrow \Delta$, and $\Gamma \Rightarrow \Delta, A$. Then, either (3) $v(\gamma) \neq \mathbf{T}$ for some $\gamma \in \Gamma$ or $v(\delta) \neq \mathbf{F}$ for some $\delta \in \Delta$ or (4) $v(B) \neq \mathbf{F}$ and $v(A) \neq \mathbf{T}$. If (3) holds, then clearly $\models_{GC4^+} A \rightarrow_w B, \Gamma \Rightarrow \Delta$. If (4) holds, then from the semantic relation of $\rightarrow_w$, it follows that $v(A \rightarrow_w B) \neq \mathbf{F}$ and again $\models_{GC4^+} A \rightarrow_w B, \Gamma \Rightarrow \Delta$.

$(\sim\wedge L)$:

$$\frac{\sim A, \Gamma \Rightarrow \Delta \quad \sim B, \Gamma \Rightarrow \Delta}{\sim (A \wedge B), \Gamma \Rightarrow \Delta}\ (\sim\wedge L)$$

Suppose that $\models_{GC4^+} \sim A, \Gamma \Rightarrow \Delta$, and $\models_{GC4^+} \sim B, \Gamma \Rightarrow \Delta$. Then, either (5) $v(\gamma) \neq \mathbf{T}$ for some $\gamma \in \Gamma$ or $v(\delta) \neq \mathbf{F}$ for some $\delta \in \Delta$ or (6) $v(\sim A) \neq \mathbf{F}$ or $v(\sim B) \neq \mathbf{F}$. If (5) holds, then clearly $\models_{GC4^+} \sim(A \wedge B),\ \Gamma \Rightarrow \Delta$. If (6) holds, then from the definition of $\wedge$, it follows that $v(A \wedge B) \neq \mathbf{T}$, whence $v(\sim(A \wedge B)) = \mathbf{T}$, and again $\models_{GC4^+} \sim(A \wedge B),\ \Gamma \Rightarrow \Delta$. ■

Lemma 13 (Soundness of GC4$^+$). If $\vdash_{GC4^+} \Gamma \Rightarrow \Delta$ is provable in GC4$^+$, then $\models_{GC4^+} \Gamma \Rightarrow \Delta$.

Proof: If the sequent $\Gamma \Rightarrow \Delta$ is an instance of axiom (ID), then $\Gamma \Rightarrow \Delta$ is valid in GC4$^+$. By induction on the depth of a derivation of $\Gamma \Rightarrow \Delta$ in GC4$^+$, it follows, by Lemma 12, that the sequent $\Gamma \Rightarrow \Delta$ is valid in GC4$^+$. ■

We are now in a position to prove the completeness of GC4$^+$. The proof below is similar to the Henkin proof described in Avron [3].

Theorem 14 (Completeness of GC4$^+$). The sequent calculus GC4$^+$ is *sound* and *complete* for $\models_{GC4^+}$.

Proof: Let us denote the provability in GC4$^+$ by $\vdash_{GC4^+}$. For any sequent Σ over the language of GC4$^+$,

$$\vdash_{GC4^+} \Sigma \text{ if } \Sigma \text{ has a proof in GC4}^+.$$

We have to prove that, for any sequent Σ over the language of GB4$^+$,

$$\models_{GC4^+} \Sigma \ \textit{iff} \ \vdash_{GC4^+} \Sigma.$$

The backward implication, representing the soundness of the system, follows immediately from Lemma 13. To prove the forward implication *completeness*, we argue by contradiction. Suppose Σ is a sequent such that $\nvdash_{GC4^+} \Sigma$. We shall prove that $\not\models_{GC4^+} \Sigma$. Let us assume that the inclusion and union of sequents are defined componetwise, i.e.,

$$(\Gamma' \Rightarrow \Delta') \subseteq (\Gamma'' \Rightarrow \Delta'') \text{ iff } \Gamma' \subseteq \Gamma'' \text{ and } \Delta' \subseteq \Delta'',$$
$$(\Gamma' \Rightarrow \Delta') \cup (\Gamma'' \Rightarrow \Delta'') = \Gamma', \Gamma'' \Rightarrow \Delta', \Delta''.$$

A sequent Σ_0 is called *saturated* if it is closed under all of the rules in GC4$^+$ applied backwards. More exactly, for any rule r in GC4$^+$ whose conclusion is contained in Σ_0, one of its premises must also be contained in Σ_0 (for a single premise rule, this means its only premise must be contained in Σ_0). For example, if $\Sigma_0 = (\Gamma_0 \Rightarrow \Delta_0)$ is saturated and $(A \rightarrow B) \in \Delta_0$, then in view of the rules $(\rightarrow R)$, we must have both $\sim A \in \Delta$ and $B \in \Delta$. In turn, if $(A \rightarrow B) \in \Gamma_0$, then in view of the rule $(\rightarrow L)$, we must have either $\sim A \in \Gamma$ or $B \in \Gamma$.

Let $\Sigma = (\Gamma \Rightarrow \Delta)$ be any sequent. We shall first prove that Σ can be extended to a saturated sequent $\Sigma^* = (\Gamma^* \Rightarrow \Delta^*)$, which is not provable in GC4$^+$. If Σ is already saturated, we are done. Otherwise, we start with the sequent Σ and expand it step by step by closing it under the subsequent rules of GC4$^+$ without losing the non-provability property. Specifically, we define a sequence $\Sigma_0, \Sigma_1, \Sigma_2, \ldots$ such that

1) $\Sigma_{i-1} \subseteq \Sigma_i$ for each $i \geq 1$,
2) Σ_i is not provable.

We take $\Sigma_0 = \Sigma_1 = \Sigma$; then, conditions 1 and 2 above are satisfied for $i = 1$. Assume that we have the constructed sequents $\Sigma_0, \Sigma_1, \ldots, \Sigma_k$ satisfying those conditions, and Σ_k is still not saturated. Then, there is a rule

$$r = \frac{\Pi_1 \ \cdots \ \Pi_l}{\Pi}$$

in GC4$^+$ such that $\Pi \subseteq \Sigma_k$ but $\Pi_i \not\subseteq \Sigma_k$ for $i = 1, \ldots, l$.

Since Σ_k is not provable, there must be an i such that $\Sigma_k \cup \Pi_i$ is not provable. Indeed, if $\Sigma_i \cup \Pi_i$ were provable for all $i, 1 \leq i \leq l$, then we could deduce $\Sigma_k \cup \Pi$ from the provable sequents $\Sigma_k \cup \Pi_i$, $i = 1, \ldots, l$, using rule r, which in view of $\Sigma_k \cup \Pi = \Sigma_k$ would contradict the fact that Σ_k is not provable. Thus, there is an $i_0, 1 \leq i_0 \leq l$, such that $\Sigma_k \cup \Pi_{i_0}$ is not provable, and we take $\Sigma_{k+1} = \Sigma_k \cup \Pi_{i_0}$. Obviously, the sequents $\Sigma_0, \Sigma_1, \ldots, \Sigma_{k+1}$ satisfy conditions 1 and 2 above.

Since all of the rules in GC4$^+$ have the subformula property, it is clear that after a finite number n of such steps, we will have added all possible premises of the rules r in GC4$^+$ whose conclusions are contained in the original sequent Σ or its descendants in the constructed sequence, obtaining a saturated extension $\Sigma^* = \Sigma_n$ of Σ, which is not provable in GC4$^+$.

Thus, we have

- $\Sigma^* = (\Gamma^* \Rightarrow \Delta^*)$ is closed under the rules in GC4$^+$ applied backwards,
- $\Gamma \subseteq \Gamma^*, \Delta \subseteq \Delta^*$,
- $\nvdash_{GC4^+} \Sigma^*$.

We use Σ^* to define a counter-valuation for Σ, i.e., a legal valuation v under the model of GC4$^+$ such that $v \not\models_{GC4^+} \Sigma$. For any propositional symbol $p \in P$ evaluated with the function defined in Definition 10, namely, we put:

$$v(p) = \begin{cases} \mathbf{T} \text{ if } p \in \Gamma \text{ and } p \notin \Delta \\ \mathbf{F} \text{ if } \sim p \in \Gamma \text{ and } \sim p \notin \Delta \\ \mathbf{B} \text{ if } \{p, \sim p\} \in \Gamma \\ \mathbf{N} \ otherwise \end{cases} \quad (1)$$

For the valuation for the strong negation in GC4$^+$, define the following:

$$v(\neg p) = \begin{cases} \mathbf{T} \text{ if } v(p) \in \{\mathbf{F}, \mathbf{N}\} \\ \mathbf{F} \text{ if } v(p) \in \{\mathbf{T}, \mathbf{B}\} \end{cases} \quad (2)$$

For any A, B of the set of all well-formed formulas of GC4$^+$,

$$v(\sim A) = \sim v(A) \quad (3)$$

$$v(A \rightarrow_w B) = \begin{cases} \mathbf{T} \text{ if } v(A) \neq \mathbf{T} \text{ or } v(B) \neq \mathbf{F} \\ \mathbf{F} \text{ if } v(A) = \mathbf{T} \text{ and } v(B) = \mathbf{F} \\ otherwise \end{cases} \quad (4)$$

It is easy to see that v defined as above is a well-defined mapping of the formulas of GC4$^+$ into **4**. Indeed, as Σ^* is not provable in GC4$^+$, then by (1), $v(p)$ is uniquely defined for any propositional symbol p, whence by (2, 3), $v(\varphi)$ is uniquely defined for any well-formed formula.

Moreover, by (2, 3), v is a legal interpretation of the language of GC4$^+$ under the interpretation of GC4$^+$, for the interpretations of $\sim, \rightarrow_w$ under v are compliant with the truth tables of those operations for this interpretation.

As Σ^* is an extension of Σ, in order to prove that $\not\models_{GC4^+} \Sigma$, it suffices to prove that $\not\models_{GC4^+} \Sigma^*$. We should prove for any well-formed formulas φ,

$$\models_{GC4^+} \gamma \text{ for any } \gamma \in \Gamma^*, \ \not\models_{GC4^+} \delta \text{ for any } \delta \in \Delta^*. \quad (5)$$

Equation (5) is proved by structural induction on the formulas in $S = \Gamma^* \cup \Delta^*$.

We begin with literals in S, having the form of either p or $\sim p$, where $p \in P$. We have the following cases:

- $\varphi = p$. Then, by (1) and the fact that Γ^* and Δ^* are disjoint (for otherwise Σ^* would be provable), we have: $v(\varphi) \neq \mathbf{F}$ if $\varphi \in \Gamma^*$ and $v(\varphi) \neq \mathbf{T}$ if $\varphi \in \Delta^*$
- $\varphi = \sim p$. If $\varphi \in \Gamma^*$, then by (1), $v(p) \neq \mathbf{T}$, whence $v(\varphi) = \sim\mathbf{F} = \mathbf{T}$ by (3). In turn, if $\varphi \in \Delta^*$, then $\varphi \notin \Gamma^*$, whence $v(p) \neq \mathbf{F}$ and $v(\varphi) = \sim v(p) \neq \mathbf{T}$.
- $\varphi = \neg p$. If $\varphi \in \Gamma^*$, then by (1) $v(p) \neq \{\mathbf{T}, \mathbf{B}\}$, whence $v(\varphi) = \mathbf{T}$ by (2). In turn, if $\varphi \in \Delta^*$, then $\varphi \notin \Gamma^*$, whence $v(p) \neq \{\mathbf{F}, \mathbf{N}\}$ and $v(\varphi) = \sim v(p) = \mathbf{F}$.

Here, we define the rank ρ of formula φ by

$$\rho(p) = 1, \rho(\sim\varphi) = \rho(\varphi) + 1, \rho(\varphi \rightarrow \psi) = \rho(\varphi) + \rho(\psi) + 1$$

Now we assume that the definition in (5) is satisfied for the formulas in S of rank up to n and suppose that $A, B \in S$ are at most of rank n. We prove that (5) holds for $\sim B$, $B \wedge C$ and $B \vee C$.

We begin with negation. Let $\varphi = \sim A$. As the case of $A = p \in P$ has already been considered, it remains to consider the following two cases:

- $A = \sim B$. Then, we have $\varphi = \sim\sim B$.
  - If $\varphi \in \Gamma^*$, then by rule $(\sim\sim L)$, we have $B \in \Gamma^*$, since Σ^* is a saturated sequent. Hence, by inductive assumption, $v(B) = \mathbf{T}$, and by (3), $v(\varphi) = \sim\sim\mathbf{T} = \mathbf{T}$.
  - In turn, if $\varphi \in \Delta^*$, then by rule $(\sim\sim R)$, we have $B \in \Delta^*$, whence by inductive assumption, $v(B) = \mathbf{F}$, and in consequence, $v(\varphi) = \sim\sim\mathbf{F} = \mathbf{F}$.
- $A = B \wedge C$. We again have two cases:
  - If $\varphi \in \Gamma^*$, then by rule $(\sim\wedge L)$, we have $\sim B, \sim C \in \Gamma^*$ since Σ^* is saturated. Hence, by inductive assumption, $v(B) \neq \mathbf{T}$ and $v(C) \neq \mathbf{T}$ (because $v(\sim B) \neq \mathbf{F}$ and $v(\sim C) \neq \mathbf{F}$). Thus, by the truth table $v(B \wedge C) \neq \mathbf{T}$; therefore, $v(\varphi) = \sim\mathbf{F} = \mathbf{T}$.
  - If $\varphi \in \Delta^*$, then by rule $(\sim\wedge R)$, we have either $\sim B \in \Delta^*$ or $\sim C \in \Delta^*$. By inductive assumption, this yields either $v(B) \neq \mathbf{T}$ or $v(C) \neq \mathbf{T}$. Thus, by the truth table, $v(B \wedge C) \neq \mathbf{T}$, whence $v(\varphi) = \sim\mathbf{T} = \mathbf{F}$.
- $A = B \vee C$. We again have two cases:
  - If $\varphi \in \Gamma^*$, then by rule $(\sim\vee L)$, we have $\sim B, \sim C \in \Gamma^*$ since Σ^* is saturated. Hence, by inductive assumption, $v(B) \neq \mathbf{T}$ and $v(C) \neq \mathbf{T}$ (because $v(\sim B) \neq \mathbf{F}$ and $v(\sim C) \neq \mathbf{F}$). Thus, $v(B \vee C) \neq \mathbf{T}$, and $v(\varphi) = \sim\mathbf{F} = \mathbf{T}$.
  - If $\varphi \in \Delta^*$, then by rule $(\sim\vee R)$ we have either $\sim B \in \Delta^*$ or $\sim C \in \Delta^*$. By inductive assumption, this yields either $v(B) \neq \mathbf{F}$ or $v(C) \neq \mathbf{F}$. Thus, $v(B \vee C) \neq \mathbf{F}$, whence $v(\varphi) \neq \mathbf{T} = \mathbf{F}$.

It remains to consider implication. Let $\varphi = A \rightarrow_w B$. We have the following two cases:

- $\varphi \in \Gamma^*$. Then, as Σ^* is saturated, by rule $(\rightarrow_w L)$, we have either $A \in \Delta^*$ or $B \in \Gamma^*$. In view of (1) and (3), and the fact that $\varphi \notin \Delta^*$, this yields either $v(A) \in$

$\{\mathbf{F}, \mathbf{N}\}$ or $v(B) \in \{\mathbf{T}, \mathbf{B}\}$. Thus $v(A \rightarrow_w B) \neq \mathbf{F}$, and $v(\varphi) = \mathbf{T}$.

- $\varphi \in \Delta^*$. Then, as Σ^* is saturated, by rules $(\rightarrow_w R)$ we have $A \in \Gamma^*$ and $B \in \Delta^*$. In view of (1) and (3), and the fact that $\varphi \notin \Gamma^*$, this yields $v(A) \in \{\mathbf{T}, \mathbf{B}\}$ and $v(B) \in \{\mathbf{F}, \mathbf{N}\}$, thus, $v(A \rightarrow_w B) \neq \mathbf{T}$, and $v(\varphi) = \mathbf{F}$.

Thus, (5) holds, and $\models_{GC4^+} \Sigma$, which ends the completeness proof. ■

GC4$^+$ may be one candidate for the extended version of decision logic that is needed to handle uncertain information and be tolerant to inconsistency.

VIII. Conclusion and Future Work

In this paper, we propose an extension of the decision logic of rough sets to handle uncertainty, ambiguity and inconsistent states in information systems based on rough sets. We investigate some properties of information system based on rough sets and define some characteristics of a certain relationship for the interpretation of truth values. We obtain some observations for a relationship between the interpretation with four-valued truth values and the regions defined with rough sets. To handle these characteristics we have introduced partial semantics with consequence relations for the axiomatization with many-valued logics and proposed a unified formulation of the decision logic of rough sets and many-valued logics. We also extend the language of many-valued logics with weak negation to enable the deduction theorem or the rule of modus ponens. We have shown that the system GC4$^+$ is sound and complete with Belnap's four-valued semantics.

In future work, the extension of language should be investigated, e.g., an operator to handle the granularity of objects or the uncertainty of a proposition, which is related to some kind of modal operators to recognize the crispness of objects. In this paper, we introduce rules of weak negation and weak implication to extend many-valued logics to handle a deduction system more usefully. To grasp the information state represented with information in detail, another extension of language should be investigated, such as modal type operators in a paraconsistent version of Łukasiewicz logic J3 [20]. Furthermore, we need to investigate another version of decision logics based on an extended version of rough set theories, e.g., the variable precision rough set (VPRS) [21]. VPRS models are an extension of rough set theory, which enables us to treat probabilistic or inconsistent information in the framework of rough sets. By these further investigations, a much more useful version of extended decision logic is expected for practical application and actual data analytics.

References

[1] Y. Nakayama, S. Akama, and T. Murai, "Deduction System for Decision Logic based on Partial Semantics," SEMAPRO 2017 The Eleventh International Conference on Advances in Semantic Processing, 2017, pp. 8–11.

[2] Z. Pawlak, "Rough Sets: Theoretical Aspects of Reasoning about Data," Kluwer Academic Publishers, 1991.

[3] A. Avron and B. Konikowska, "Rough Sets and 3-Valued Logics," Studia Logica, vol. 90, 2008, pp. 69–92.

[4] D. Ciucci and D. Dubois, "Three-Valued Logics, Uncertainty Management and Rough Sets," in Transactions on Rough Sets XVII, Lecture Notes in Computer Science book series (LNCS, volume 8375), 2001, pp. 1–32.

[5] A. Urquhart, "Basic Many-Valued Logic," Handbook of Philosophical Logic, vol. 2, 2001, pp. 249–295.

[6] T.-F. Fan, W.-C. Hu, and C.-J. Liau, "Decision logics for knowledge representation in data mining," in 25th Annual International Computer Software and Applications Conference. COMPSAC, 2001, pp. 626–631.

[7] Y. Lin and L. Qing, "A Logical Method of Formalization for Granular Computing," IEEE International Conference on Granular Computing (GRC 2007), 2007, pp. 22–22.

[8] S. Akama, T. Murai, and Y. Kudo, "Reasoning with Rough Sets, Logical Approaches to Granularity-Based Framework," 2018.

[9] J. Nuel D. Belnap, "A Useful Four-Valued Logic," in Modern Uses of Multiple-Valued Logic, vol. 2. Reidel Publishing, 1977, pp. 5–37.

[10] A. R. Anderson and J. Nuel D. Belnap, "Tautological Entailments," Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, vol. 13, 1975, pp. 9–24.

[11] J. Van Benthem, "Partiality and Nonmonotonicity in Classical Logic," Logique et Analyse, vol. 29, 1986, pp. 225–247.

[12] R. Smullyan, "First-Order Logic," Dover Books, 1995.

[13] R. Muskens, "On Partial and Paraconsistent Logics," Notre Dame J. Formal Logic, vol. 40, 1999, pp. 352–374.

[14] V. Degauquier, "Partial and paraconsistent three-valued logics," Logic and Logical Philosophy, vol. 25, 2016, pp. 143–171.

[15] S. Akama and Y. Nakayama, "Consequence relations in DRT," Proc. of The 15th International Conference on Computational Linguistics COLING 1994, vol. 2, 1994, pp. 1114–1117.

[16] G. Priest, "The Logic of Paradox," Journal of Philosophical Logic, vol. 8, 1979, pp. 219–241.

[17] S. Kleene, "Introduction to Meta-mathematics," 1952.

[18] G. Priest, "An Introduction to Non-Classical Logic From If to Is 2nd Edition," 2008.

[19] P. Doherty, "NM3 - A three-valued cumulative non-monotonic formalism," in Logics in AI, European Workshop (JELIA), 1990, pp. 196–211.

[20] R. L. Epstein, "The Semantic Foundations of Logic," 1990.

[21] W. Ziarko, "Variable precision rough set model," Journal of Computer and System Science 46, 1993, pp. 39–59.

Pro2Screen – High-Fidelity-Prototyping of Mobile Enterprise Applications using Process Models

Leo Fuhr, Matthias Jurisch, Stephan Böhm, Tobias Braumann, Bodo Igler
Faculty of Design – Computer Science – Media
RheinMain University of Applied Sciences
Wiesbaden, Germany
Email: leo.fuhr@hs-rm.de, matthias.jurisch@hs-rm.de, stephan.boehm@hs-rm.de, tobiasbraumann@gmail.com, bodo.igler@hs-rm.de

***Abstract*—Mobile Enterprise Applications are becoming more and more relevant to enterprises as the dissemination of smartphones has risen over the last decade. However, developing these applications is a very challenging and resource-intensive task. In this context, prototyping can lead to several benefits, including a better app usability. While Mobile Enterprise Applications are often used to support or carry out business processes, no low-code mobile prototyping approach exists that is based on business process models and is adapted to the needs of non-developers. In this paper, we present the tool Pro2Screen that fills this gap. The tool uses a Business Process Model and Notation (BPMN) model annotated with screen designs as a source for generating a prototype. The prototype is integrated with a business process execution engine that runs the business process. To demonstrate the effectiveness of Pro2Screen, we present a user study with experts from the field of Mobile Enterprise application development. The user study shows that the approach is able to tackle important aspects of prototyping in Mobile Enterprise Application development and enables practitioners with little or no coding experiences to develop prototypes that closely resemble functional aspects of a final product.**

***Keywords*–Mobile Enterprise Application; BPMN; Process Model; Prototyping.**

I. Introduction

Since the beginning of the decade, mobile applications have become more and more ubiquitous. This trend also reached enterprises, where employees expect to use smartphone apps for their daily work with the high usability they are accustomed to from using consumer apps. These expectations and the continuously and fast changing ecosystem of mobile app development pose a significant difficulty for the development of Mobile Enterprise Applications (MEA). In our previous work, we proposed the prototyping approach Pro2Screen for MEAs [1] that is extended in this paper.

MEAs differ from regular consumer apps in several ways, e.g., they are often used to support some kind of business process, have only few potential users in comparison to consumer apps and need to adhere to enterprise specific guidelines [2]. Integrating business processes into mobile applications means implementing new interfaces to process engines and adhering to process guidelines are some of the challenges for MEA development that are caused by MEA-specific aspects. Since other aspects of mobile application development also need to be taken care of, these factors contribute to developing MEAs being a time consuming and expensive process.

To reduce the effort required to develop mobile applications in general, prototyping can be used. A good prototyping process can prevent misunderstandings and make the conceptual phase of the development process prior to coding significantly easier to handle and therefore reduce costs [3][4]. More important, this can also allow a better usability of the final product which will improve the willingness of employees to use the final MEA. However, no prototyping tool that supports all of the aforementioned aspects of MEAs exists. To our knowledge, there is no prototyping approach that caters to the business process aspect of this type of application and does not require coding.

This paper is an extension of [1], which presented the prototyping approach Pro2Screen. Pro2Screen focuses on using business process models [5] as the primary source for MEA prototypes. Our approach enables designers and business engineers to create prototypes with several user roles and process steps that connect these roles. For the creation of the prototype, no coding experience or background in formal modeling is required. This is accomplished by annotating visual representations of business process models with screen designs and creating a prototype using code generation and business process execution engines that can interpret and automatically execute business processes.

The description of the approach is extended by a discussion of an evaluation with MEA development experts and a discussion of the results of the approach. The evaluation consists of a user study, where users were given a task to fulfill using Pro2Screen and answer a survey. This work is embedded into the scope of the Prototyping Framework for Mobile App Design in Large Enterprises (PROFRAME) [6]. The presented work will lay the foundation for the implementation of PROFRAME.

The remainder of this paper is structured as follows: Section II gives a brief overview on related work and identifies the research gap. The general approach of this paper is presented in Section III. Details on the behavioural modeling of the prototypes are given in Section III-A, designing screens is discussed in Section III-B and code generation and prototype execution are presented in Section III-C. The implementation is described in Section III-D. A user study and its results are presented in Section IV. The methodology of the evaluation is discussed in Section IV-A. Section IV-B presents the questionnaire. Results of the questionnaire are discussed in Section IV-C. Section V discusses advantages

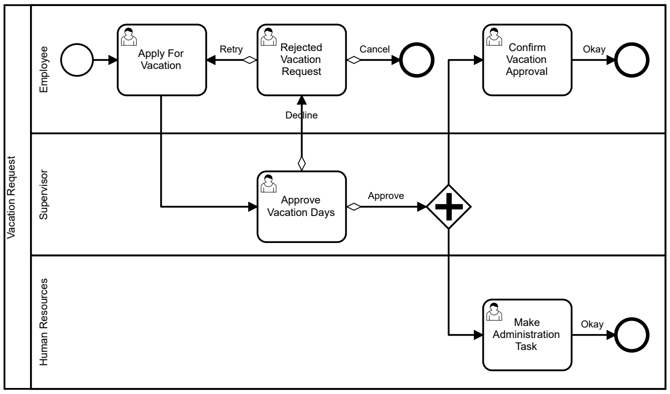


Figure 1. Example Process

and disadvantages of Pro2Screen including findings from the evaluation. A conclusion is given in Section VI.

II. Related Work

According to [2], a huge gap between the development of MEAs and standard non-mobile enterprise applications can be observed. However, the demand for MEA development in the next few years will be much higher than the supply [7]. Hence, it is important to support a very efficient way of implementing MEAs.

One way to improve the development of MEAs is improving the prototyping process. Several models for classifying prototypes have been proposed in the literature. Nielsen [4] proposed a distinction between vertical and horizontal prototyping fidelity. A horizontal prototype supports most functionalities of a product, whereas a vertical prototype allows only a few functionalities but is technically more similar to the final product. The filter fidelity model [8] adds more dimensions to this view, e.g., regarding interactivity, data model, weight and many other dimensions. Breadth and depth of functionality are also included in this model.

For prototyping mobile applications in general, many products and approaches can be found in the literature. However, regarding prototyping for MEAs, only a few tools can be found (e.g., Kony, Verivo Akula and SAP Mobile) [9]. These tools are often focused on a specific use case or bound to a specific platform. None of them take business process modeling into account, so the depth of functionality according to the filter fidelity model is low.

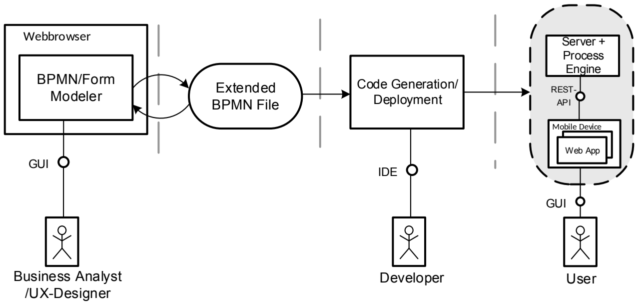


Figure 2. Pro2Screen Architecture

Integrating process models into application development has been discussed in the area of process-driven development. AgilePDD [10] proposes an agile approach to implementing business processes. In the prototyping phase of this process, business process models are used to define the behaviour of the prototype. While this approach seems promising, it does not define how a prototype should be generated from the process model or how process steps should be represented as screens. The approach is in general focused on modeling use cases with business process models, whereas generating code from these models is only mentioned as an option to be considered [11].

Similar work has been conducted in the area of model-driven development using models formulated in the Business Process Model and Notation (BPMN). BPMN [12] is the most popular language for modeling business processes [13][14]. This area focusses on supporting the coding of applications by generating parts of applications from BPMN models. A platform used in this context is WebRatio [15]. WebRatio allows the implementation of mobile applications by modeling them in BPMN. While WebRatio can be used for prototyping, the platform tries to help developers with a background in coding and formal modeling and can not be used by users not experienced with these techniques. Other approaches like Axon.ivy [16] is a similar approach that is even more tailored at coders.

From the presented literature, we can conclude that no prototyping approach or tool for MEAs exists that facilitate a high fidelity regarding the representation and integration of business process models into the prototype and cater to the needs of UX designers and Business Engineers that do not have a background in coding. Also, there should be no requirement for a deep understanding of formal modeling approaches prior to using the tool. This issue is at the core of our research, since a prototype that better resembles the final product can help improve its usability.

III. Approach

We consider three major requirements for our work: the tool needs to support (1) modeling a business process, (2) designing a user interface and (3) generating a platform-independent prototype that can be executed on mobile devices. The basic idea of our approach is to use business process models as the primary source for the prototype. The process model defines the behaviour of the app. To add a graphical user interface, the process model is annotated with screen designs for specific parts of the business process. With this information, a prototype for the app is generated.

In practice, several implementations of process engines exist. They are able to interpret and execute BPMN models and integrate business processes with several backend systems. Therefore, BPMN is used as the process model language in the presented prototyping tool.

An overview of the prototyping process is given in Figure 2. A Business Analyst/UX Designer uses the *BPMN/Form Modeler* to create a model of the process that shall be implemented with an app including a user interface design. To model the process itself, one can simply use an existing process modeling tool that supports BPMN. This can be done in close coordination with the customer, e.g., at a kick-off-meeting for a project. The result of this process is an *extended BPMN file*. Internally, this file can then be used to generate a

Web App that cooperates with a business *process engine*. The customer can then use this app as a prototype, which allows a clear separation of the code generation and the prototype modeling. For customization, a developer that modifies the code generation can be included in the process.

To support the described process, answers to the following questions are required: What aspects of a process should be represented as screens (Section III-A)? How can screens be designed and how can data be reused over several screens (Section III-B)? How are prototypes generated (Section III-C)?

A. Process Model

BPMN in general is well-known for its graphical representation of business processes. An example model is shown in Figure 1. The most important element of BPMN is the *task* (e.g., *Apply for Vacation*). Tasks represent any kind of activity. Several kinds of tasks exist, the kind of task is represented by a icon at the top of a task. *Apply for vacation* is a user task and *Check Vacation request* is a script task. User tasks require user interaction whereas script tasks are automatically executed by the business process engine.

To connect tasks, so called *Sequence Flows* that are represented by arrows are used. Gateways (represented by rhombuses) are used to model situations where the flow is split, either because of decisions (x) or parallel execution (+). The swimming pool element (*Vacation Request*) is used to structure the control flow. A swimming pool can contain multiple swimlanes (e.g., *Employee*) that distinguish different domains of activity.

Our approach proposes a representation of tasks as screens: when the model is executed, each user task corresponds to one screen on the mobile device. A swimming pool corresponds to an app and a swimlane corresponds to a user role. For the example shown in Figure 1, users with role *Employee* would be shown at most three different screens (*Apply for Vacation, Vacation Request Rejected* and *Confirm Vacation Approval*) and users with role *Supervisor* or *HR* one (*Approve Vacation Days* and *Start Administrative Task*). Sequence flows determine the control flow of the business process.

B. Form Modeler

To design the forms that correspond to user tasks, a form modeler is used that is able to store screen designs as annotations in BPMN files. The form modeler needs to add a screen design to each user task and store the design as an annotation in the BPMN model. A screenshot of the form modeler that implements this idea is shown in Figure 3. The user of the form modeler can drag and drop user interface components (1), e.g., *Plain Text*, *Text Inputs* and *Radio Buttons*, into the screen layout (2). Properties of components can be modified using controls on the right (3).

Our approach uses a grid layout to model the screen design. By using a grid layout, the prototype is not bound to a specific screen size or orientation. The grid is shown as dashed lines in the screenshot. Users can add and remove rows and columns. Each cell in this grid can only hold one widget. To improve the design, the user can modify row height and column width.

By modifying a component's properties using the box on the right (3), the user can edit several aspects regarding its behaviour and appearance, e.g., inputs can be set as editable and required and their label can be defined. The property *parameter* (4) is used to specify parameter IDs that are used to identify data throughout the complete business process. When a screen is used to input data into a field with a certain parameter ID and another screen shown later in the process has a component with a matching parameter ID, the second screen will show the data entered in the first screen. The parameter IDs are identifiers in a global data space bound to a workflow.

The screen shown in the example corresponds to the task *Apply for Vacation* from Figure 1. To view the data entered in this screen, e.g., in the task *Approve Vacation Days*, it is only required to add an UI component to that task and set its parameter ID to `request_reason`, similar to the example shown in Figure 3 (4).

C. Code Generation and Process Execution

The previously described steps allow the creation of an annotated BPMN model that contains information about the behaviour of the application, as well as the UI design. Based on this information, code generation can be used to create a prototype.

Besides generating app prototypes, using BPMN as a foundation for the prototype allows execution of the process model on a business process execution engine. To exploit this circumstance, the generated prototype is separated into two parts: (1) A business process engine that is given the business process model and executes it and (2) a Web App that interacts with the business process execution engine. The engine controls the process and data related to it. This allows the synchronization between prototypes for different user roles involved in the process, which are all created in the generation process. Also, the business process engine can be integrated with other enterprise systems, which allows accessing real-world data from the prototype.

D. Implementation

As a component for modeling business processes, Camunda Modeler [17] is used. Camunda Modeler needed to be extended to provide an interface to the form modeler. Angular [18] is used to implement the form modeler from scratch to allow a seamless integration with the process modeler. These components write their data into the shared extended BPMN file.

Prototype generation form the shared BPMN document is implemented using XSLT [19]. To support multiple mobile platforms, the generated code uses the Ionic framework [20], a HTML and Javascript-based Framework, as SDK. Ionic allows the visual design of the app to be easily changed using CSS. This supports the integration of enterprise corporate design guidelines into the product. To execute the business process, the Camunda Core Engine [21] is used. To interact with the engine, the Web App uses the Camunda REST API.

IV. Evaluation

The core of the Pro2Screen idea is to support Business Analysts and UX Designers to develop prototypes in close coordination with the customers. We want to evaluate the suitability of this concept in a user study with the help of a group of practitioners who handle these tasks in their daily work. This evaluation shall answer the following research questions:

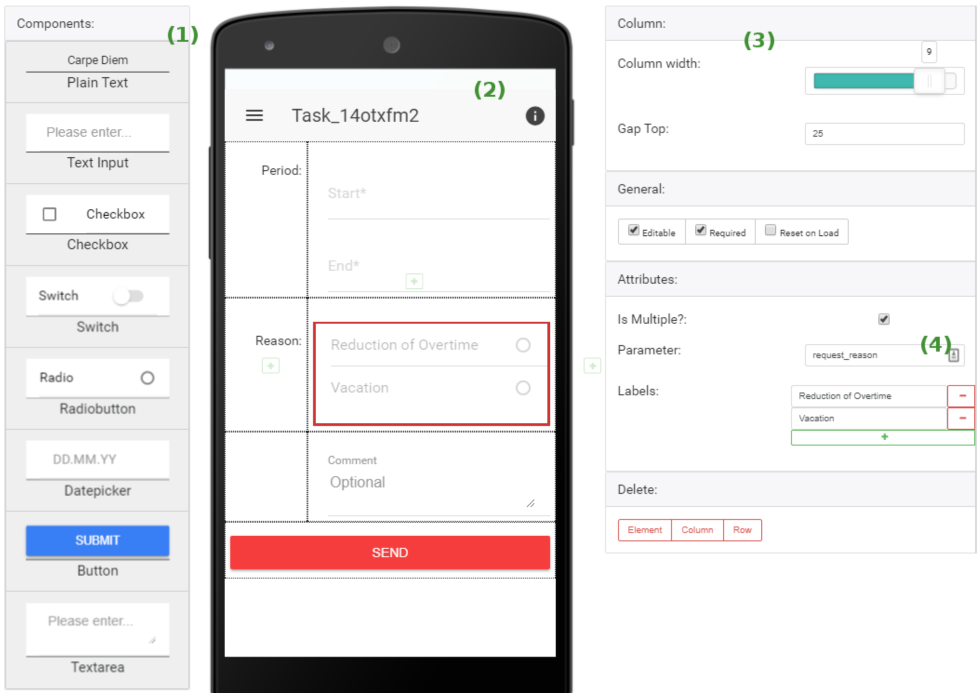


Figure 3. Form Modeler Screenshot

- Can Business Analysts and UX Designers that have no experience in coding mobile apps develop prototypes using Pro2Screen?
- Do practitioners see a benefit in working with the tool?
- Is experience with BPMN-based modeling tools required to work with Pro2Screen?

As described in Section III Pro2Screen consists of two parts, the BPMN-modeler and the form modeler. In our experiment, the group of practitioners worked on two subtasks: First, model a business process using the BPMN-modeler and second, model the corresponding interactive screens for the app using the form modeler. Test participants filled out questionnaires before and after completing the two subtasks. In the following subsections the methodology, the questionnaire and the results are described.

A. Methodology

The evaluation is carried out on two days with a total of seven test persons. The test persons are employees of a leading German ICT solution provider and bring along experiences from their daily work in the process of Mobile Enterprise Application Development from requirement analysis (Business Engineers) or User Experience Design (UX Designers and Consultants).

In total, a test run takes 90 minutes. Each test run is conducted in a separate room and monitored by two test supervisors. The test runs are carried out sequentially under the same conditions. First, the concept of Pro2Screen is presented by the supervisors. The aim of the testing is explained, too. Subsequently, the test persons receive a questionnaire about their personal experiences in the field of MEA development. The experience with prototyping tools as well as the modeling of a business process in BPMN are also part of this questionnaire.

Next, the test persons watch a three-part video tutorial as an introduction to the functionality and modeling capabilities of Pro2Screen. In this tutorial, the business process for a vacation request from Section III-A is described in more detail. In the first part of the tutorial the final BPMN process model with the corresponding screens is shown. How the BPMN process can be modeled in Pro2Screen is shown in the second part of the tutorial. The last part of the video shows how to model the corresponding screens in the form modeler.

Afterwards, the test persons fill out a pre-test questionnaire with seven questions before they could use the tool independently on the basis of two tasks: First, sketch a business process in the BPMN modeler and second, model the screens with the help of the form modeler. As an input for the task to be completed, the following use case description is given to the test participants: *A procurement process involves the roles of employee, manager and a member of the purchasing department. The need for procurement is recognized and initiated by the employee himself. An employee can choose between a laptop, PC, or monitor. After successful selection, this process is transmitted to the manager. The manager can reject or approve the acquisition request of the employee. If rejected, the process is sent back to the employee without explanation. He takes note of this and ends the process. If the acquisition request is approved by the manager, once the employee receives a positive response regarding his application and on the other hand, the purchasing department on the acquisition request are informed. Both roles end the assigned subprocess.* The participants have 10 minutes to sketch this process using the BPMN-modeler and 20 minutes to add forms to the process using the screen modeler. After solving the two tasks, the post-test questionnaire is distributed. The same questions are asked on the post-test questionnaire as on the pre-test questionnaire.

B. Questionnaire/Survey

In this Section, the contents of the questionnaires are presented in detail. During the evaluation three questionnaires are distributed. The user questionnaire is used to gather information regarding the user and experiences with BPMN and prototyping and opinions on MEA prototyping in general. A pre-test concept evaluation questionnaire and a post-test questionnaire try to capture the users opinions and proficiency of Pro2Screen before and after the users execute the task.

The experience values in the user questionnaire relate to the methods used in this tool. Each question can be answered on a scale of 1 (very low or never) to 5 (very high or always). The questions are the following:

1) *UX proficiency*: How do you rate your skills in UX Design?
2) *Prototyping tool usage*: How many times have you worked with a mobile app prototyping tool?
3) *BPMN proficiency*: How do you assess your business process modeling skills?
4) *BPMN tool usage*: How many times have you worked with a BPMN modeling tool?

Additionally, each user rated the following statements on a scale of 1 (do not agree at all) to 5 (completely agree).

1) *Existing tools sufficiency*: Existing app development tools are sufficient.
2) *Enjoy experimenting with tools*: I enjoy experimenting with tools that support app development.
3) *Large enterprise adaption*: Existing tools are well adapted to being used in large enterprises.
4) *Collaboration support*: Existing tools support collaboration in interdisciplinary teams.
5) *Reuse support*: Existing tools support the reuse of existing results.

The questions asked in the pre- and post-test regarding Pro2Screen are identical. Each question can be answered on a scale of 1 (very bad) to 5 (very good).

1) *Conceptual comprehension*: How much do you think you have understood the concept of the new tool so far?
2) *Concept rating*: How do you rate the concept of this new tool so far?
3) *Enterprise-wide applicability*: How do you rate the usability of this tool in your company so far?
4) *Local applicability*: How do you rate the applicability of this tool in your personal workspace?
5) *Self-assessed proficiency*: How do you think so far, will you be able to cope with this tool?
6) *Personal job ease*: How do you rate the chances that this novel tool could make your job easier?
7) *App quality improvement*: How do you rate the chances that this tool will improve the quality of app development?

C. Results

We executed the aforementioned test plan with seven users from a large enterprise that work on mobile enterprise applications. The created solutions of the test persons were compared with the model solution. These results are presented in Table I. In the first task they should create the BPMN process diagram for the described task from Section IV-A. At this task a maximum of 18 points could be achieved. In each case, one point was awarded for creating one of the three roles in the swimming lane, setting the correct task for the corresponding roles and setting the associated connections between the tasks. On average, each test person scored 12.86 points, which corresponds to 71.43% of the possible points.

In the second task the test persons should model the screens of the BPMN process diagram described in Section IV-A. Due to the comparison of the results, they start with the same BPMN process diagram. This diagram is the model solution of the first task. In the second task a maximum of 24 points could be achieved. They could receive one point for setting the correct fields. Further points could be reached if the necessary fields were linked to global variables and the buttons on the screens were linked to the corresponding outgoing transitions. The test persons achieved an average of 19.5 points in this task. This corresponds to 78% of total points. In sum, for both tasks, the test persons reached an average of 32.36 points out of 42 possible points. In other words, 75.25% of all possible points were achieved by the test persons on average.

TABLE I. Completion Rate

	Task 1	Task 2	Total
Average number of achieved points	12.86	19.50	32.36
Percentage of achieved points	71.43%	78.00%	75.25%

An overview of their answers to the user questionnaire are shown in Table II. In these results, we can see that the majority of the test persons are experienced in UX Design. But on the other hand, the majority has low experience (the median is one, meaning no experience using BPMN-based tools) modeling business processes with a BPMN modeling tool. While the capability of modeling business processes in general is rated a little higher than working with BPMN tools, the values are quite spread out. The answers to this question is quite similar to the regular usage of prototyping tools, while test persons seem to use these tools a lot more than business process modeling tools. Nearly all test persons enjoy experimenting with tools.

Regarding the experts opinions on existing MEA development tools, the majority thinks that prototyping tools are not well adapted to large enterprises. The users did not express that existing tools are sufficient for MEA development, while they are not completely unusable for this task. Similar answers were given regarding the support for collaboration or reuse.

An overview of the pre- and post-test evaluation of these users is shown in Table III. In nearly all questions, the scores slightly improve from pre-test to post-test. The results for *conceptual comprehension* are at least four in pre- and post-test. This means that after watching the video tutorial (pre-test) and after doing the tasks (post-test), the test persons understand the concept of the tool well. In the post-tests, these values are even higher. Results for *concept rating* are similar. The

TABLE II. Results of User Questionnaire

	Minimum	Median	Maximum
UX proficiency	2	4	5
Prototyping tool usage	1	3	4
BPMN proficiency	1	2	4
BPMN tool usage	1	1	4
Existing tools sufficiency	2	3	4
Enjoy experimenting with tools	3	4	4
Large enterprise adaption	2	2	4
Collaboration support	2	3	4
Reuse support	2	3	4

results for *enterprise-wide applicability* show that it is not clear whether the tool can be used enterprise-wide. Regarding the *local applicability*, we can see disagreement in the results. While the majority of the responses to this question in the post-test where four or higher, the results are spread out. Post-test answers to *personal job ease* show that a majority of test persons believe that Pro2Screen can benefit their personal job, with a slight increase from the pre-test. Results for a possible improvement in app quality are similar.

V. Discussion

In comparison to prototyping approaches mentioned in Section II, we see several benefits. One important advantage of using BPMN as the foundation for prototyping is that it supports reusing existing process models to create a prototype. Even an automated transformation from existing files is supported. This is not possible for prototyping approaches based on other models. Another important aspect of this approach is the option to build applications using more than one user role easily. Supporting a business process execution engine allows the integration of existing enterprise systems in the prototype, since these systems can be integrated into the process engine. This can allow the prototype to access real-world data, which give the user of the prototype a better understanding of the functionality.

A drawback of our approach is the limitation regarding visual design choices of the form modeler caused by the grid layout and the limited set of standard components. While inexperienced users might see the simplicity as an advantage, especially designers might need more freedom in positioning components and a broader collection of usable widgets.

Regarding the prototype's fidelity according to the filter fidelity model [8], our approach allows building prototypes that have a high breadth and depth of functionality and a close relation to data and appearance of the final product. This can make it easier to demonstrate to customers how an app can support their business processes and help manage expectations. This can lead to reduced costs for reworking requirements and app concepts during the MEA development process and improve the usability of the final product.

From our user study, we can see that there is room for improvement in the area of MEA-Prototyping in general. Users in our evaluation stated that existing tools are not well adapted to the needs of MEA development and did not state that existing tools are sufficient for this process. Since the majority of test persons see an applicability of Pro2Screen in their personal work area, this can indicate that this tool can make a contribution in MEA development. This effect intensified after the test users worked with the tool. Users did also see that MEA quality could be improved by working with the tool.

Our evaluation also shows that the tool is well adapted to the needs of UX Designers and Business Engineers. While the majority of test users were not familiar with prototyping tools or BPMN modeling, after the test most users believed that they can work with the tool. Even before the test, the users were confident that they can work with it. This shows that the tool can be used by users with little experience in BPMN modeling or even prototyping tools, which is especially important when prototypes are created by users that are not as ”tech-savvy” as coders.

The presented evaluation has some limitations: While the size of seven participants is in general considered suitable for usability testing [22], the ample size is small, and thus the generalizability of our results to MEA development experts is questionable. Besides, our evaluation was performed with test users from only one company. The situation in other large enterprises might be fundamentally different and thus needs to be further evaluated.

In general, the presented approach allows creating prototypes by directly mapping screens and process steps. This supports non-developers in creating high-fidelity prototypes without the need for learning to code or a background in formal modeling. Our evaluation indicates that the approach is adapted to the needs of this group, which makes the approach an interesting candidate for further research.

VI. Conclusion and Future Work

In this paper, we have presented the prototyping tool Pro2Screen for MEAs that is based on the usage of business process models written in BPMN. Prototypes are created using a business process annotated with screen designs. The annotated process is then used to generate a prototype that consists of an app and a business process execution engine that executes the process.

Pro2Screen allows fast prototyping of MEAs, since it is possible to reuse existing BPMN models for prototyping and integration with other enterprise applications through the business process execution engine. Generated prototypes can achieve a high level of fidelity regarding several aspects, especially the depth of functionality and the visual quality of the prototype is high.

To evaluate the feasibility of this approach, we have presented a user study conducted with practitioners from the field of MEA development with a background in user experience and business engineering. The evaluation shows that users from our test group where able to work with Pro2Screen and indicates that Pro2Screen can make a significant contribution in the area of MEA development.

As future work, we plan to further evaluate the benefits of this prototyping approach regarding the ability to develop MEAs. Especially financial aspects of the approach and possible improvements of MEA quality need to be examined. As shown in the discussion, to obtain generalizable results, a repetition of our evaluation with a larger and more diverse user group will also be considered. Another next step could be integrating existing screens from a standard screen library into the prototyping tool.

TABLE III. Comparison Pre-Test versus Post-Test

	Pre-Test			Post-Test		
	Minimum	Median	Maximum	Minimum	Median	Maximum
Conceptual comprehension	4	4	5	4	5	5
Concept rating	3	4	5	4	4	5
Enterprise-wide applicability	3	3	5	3	3	4
Local applicability	2	3	5	2	4	5
Self-assessed proficiency	3	4	4	3	5	5
Personal job ease	3	3	5	3	4	4
App quality improvement	3	3	4	2	4	4

ACKNOWLEDGMENT

This work was funded by the German Federal Ministry of Education and Research, grant no. 03FH032PX5; the PROFRAME project at RheinMain University of Applied Sciences. All responsibility for the content of this paper lies with the authors.

REFERENCES

[1] T. Braumann, D. Broll, M. Jurisch, and B. Igler, "Business Process Models as a Foundation for High-Fidelity-Prototypes of Mobile Enterprise Applications," CENTRIC 2017 : The Tenth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, 2017, pp. 32–35.

[2] A. Giessmann, K. Stanoevska-Slabeva, and B. de Visser, "Mobile enterprise applications–current state and future directions," in System Science (HICSS), 2012 45th Hawaii International Conference on, Jan 2012, pp. 1363–1372.

[3] F. Kordon and Luqi, "An introduction to rapid system prototyping," IEEE Transactions on Software Engineering, vol. 28, 2002, pp. 817–821.

[4] J. Nielsen, Usability Engineering. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[5] R. S. Aguilar-Saven, "Business process modelling: Review and framework," International Journal of Production Economics, vol. 90, no. 2, 2004, pp. 129–149.

[6] M. Jurisch, B. Igler, and S. Böhm, "PROFRAME: A Prototyping Framework for Mobile Enterprise Applications," in CENTRIC 2016, The Ninth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, 2016, pp. 7–10.

[7] Gartner, "Gartner says demand for enterprise mobile apps will outstrip available development capacity five to one," https://www.gartner.com/newsroom/id/3076817, website [retrieved: February, 2018], 2015.

[8] T. Hochreuter, K. Kohler, and M. Mareen, "Towards a more conscious use of prototypes in mobile user experience design," in Prototyping to Support the Interaction Designing in Mobile Application Development (PID-MAD), 2013.

[9] J. Wong, V. L. Baker, A. Leow, and M. Resnick, "Magic quadrant for mobile app development platforms," https://www.gartner.com/doc/3744122/magic-quadrant-mobile-app-development, website. [retrieved: February, 2018], 2015.

[10] A. Herden, P. Farias, P. de Andrade, and A. Albuquerque, "Agile PDD: One approach to software development using BPMN," in 11th International Conference Applied Computing, Porto, Portugal, 2014.

[11] A. Herden, P. P. M. Farias, and A. B. Albuquerque, "An approach based on BPMN to detail use cases," in New Trends in Networking, Computing, E-learning, Systems Sciences, and Engineering. Springer, 2015, pp. 537–544.

[12] Object Management Group, "Business process model and notation, version 2.0," http://www.omg.org/spec/BPMN/2.0/PDF/, website. [retrieved: February, 2018], 2011.

[13] P. Harmon, "The state of business process management 2016," http://www.bptrends.com/bpt/wp-content/uploads/2015-BPT-Survey-Report.pdf, technical report. [retrieved: February, 2018], 2016.

[14] M. Brambilla, S. Butti, and P. Fraternali, "WebRatio BPM: A Tool for Designing and Deploying Business Processes on the Web," in Web Engineering, B. Benatallah, F. Casati, G. Kappel, and G. Rossi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 415–429.

[15] WebRatio, "Webratio bpm platform," https://www.webratio.com/site/content/en/bpm-application-development, website. [retrieved: February, 2018], 2018.

[16] Axon.ivy, "digitalize your business — AXONIVY," http://www.axonivy.com/, website. [retrieved: February, 2018], 2018.

[17] Camunda, "Modeler - Camunda BPM," https://camunda.org/features/modeler/, website. [retrieved: February, 2018], 2018.

[18] Google, "Angular," https://angular.io, website. [retrieved: February, 2018], 2018.

[19] W3C, "Xsl transformations (xslt) version 1.0," https://www.w3.org/TR/xslt, website. [retrieved: February, 2018], 1999.

[20] Drifty Co., "Build amazing native apps and progressive web apps with ionic framework and angular," https://ionicframework.com/, website. [retrieved: February, 2018], 2018.

[21] Camunda, "Engine - Camunda BPM," https://camunda.org/features/engine/, website. [retrieved: February, 2018], 2018.

[22] J. Nielsen, "How many test users in a usability study?" https://www.nngroup.com/articles/how-many-test-users/, website. [retrieved: May, 2018], 2012.