IARIA

ICCGI 2015

The Tenth International Multi-Conference on Computing in the Global Information Technology

ISBN: 978-1-61208-432-9

October 11 - 16, 2015

St. Julians, Malta

ICCGI 2015 Editors
Hermann Kaindl, Vienna University of Technology, Vienna, Austria
György Kálmán, mnemonic AS, Norway
Dan Tamir, Texas State University, USA

ICCGI 2015

Forward

The Tenth International Multi-Conference on Computing in the Global Information Technology (ICCGI 2015), held between October 11 - 16, 2015 - St. Julians, Malta, comprised a series of independent tracks that complement the challenges on various facets of computation, systems solutions, knowledge processing, system implementation, and communications and networking technologies.

To topics are covering a large spectrum of topics related to global knowledge concerning computation, technologies, mechanisms, cognitive patterns, thinking, communications, user-centric approaches, nanotechnologies, and advanced networking and systems. The conference topics focused on challenging aspects in the next generation of information technology and communications related to the computing paradigms (mobile computing, database computing, GRID computing, multi-agent computing, autonomic computing, evolutionary computation) and communication and networking and telecommunications technologies (mobility, networking, bio-technologies, autonomous systems, image processing, Internet and web technologies), towards secure, self-defendable, autonomous, privacy-safe, and context-aware scalable systems.

The conference had the following tracks:

- Networking technologies
- Digital information processing
- Modeling
- Software development and deployment
- Mobile and intelligent techniques
- Optimization and performance

Similar to the previous edition, this event attracted excellent contributions from all over the world. We were very pleased to receive top quality contributions.

We take here the opportunity to warmly thank all the members of the ICCGI 2015 technical program committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to ICCGI 2015. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the ICCGI 2015 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope ICCGI 2015 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the field of computing in the global information technology. We also hope that St. Julians, Malta provided a pleasant environment during the conference and everyone saved some time to enjoy the beauty of the city.

ICCGI 2015 Chairs

CCGI Advisory Committee
Constantin Paleologu, University Politehnica of Bucharest, Romania
Tibor Gyires, Illinois State University, USA
Luc Vouligny, Institut de Recherche d'Hydro-Québec - Varennes, Canada
John Terzakis, Intel, USA
Yasushi Kambayashi, Nippon Institute of Technology, Japan
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Teemu Kanstrén, VTT Technical Research Centre of Finland, Finland
Mansour Zand, University of Nebraska, USA
Arno Leist, Massey University, New Zealand
Jean-Denis Mathias, IRSTEA, France
Dominic Girardi, RISC Software GmbH, Austria

ICCGI Special Area Chairs

Knowledge/Cognition
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Tadeusz Pankowski, Poznan University of Technology, Poland
e-Learning/Mobility
José Rouillard, Université Lille Nord, France
Industrial Systems
Beata Czarnacka-Chrobot, Warsaw School of Economics, Poland
Bernhard Freudenthaler, Software Competence Center Hagenberg GmbH, Austria

ICCGI Publicity Chair
Marek Opuszko, Friedrich-Schiller-University of Jena, Germany

ICCGI 2015

Committee

ICCGI 2015 Advisory Committee

Constantin Paleologu, University Politehnica of Bucharest, Romania
Tibor Gyires, Illinois State University, USA
Luc Vouligny, Institut de Recherche d'Hydro-Québec - Varennes, Canada
John Terzakis, Intel, USA
Yasushi Kambayashi, Nippon Institute of Technology, Japan
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Teemu Kanstrén, VTT Technical Research Centre of Finland, Finland
Mansour Zand, University of Nebraska, USA
Arno Leist, Massey University, New Zealand
Jean-Denis Mathias, IRSTEA, France
Dominic Girardi, RISC Software GmbH, Austria

ICCGI 2015 Special Area Chairs

Knowledge/Cognition
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Tadeusz Pankowski, Poznan University of Technology, Poland

e-Learning/Mobility
José Rouillard, Université Lille Nord, France

Industrial Systems
Beata Czarnacka-Chrobot, Warsaw School of Economics, Poland
Bernhard Freudenthaler, Software Competence Center Hagenberg GmbH, Austria

ICCGI 2015 Publicity Chair

Marek Opuszko, Friedrich-Schiller-University of Jena, Germany

ICCGI 2015 Technical Program Committee

Pablo Adasme, Universidad de Santiago de Chile, Chile
El-Houssaine Aghezzaf, Gent University, Belgium
Johan Akerberg, ABB Corporate Research, Sweden
Nadine Akkari, King Abdulaziz University, Kingdom of Saudi Arabia

David Chen, University of Bordeaux, France
Tzung-Shi Chen, National University of Tainan, Taiwan
Wen-Shiung Chen (陳文雄), National Chi Nan University, Taiwan
Zhixiong Chen, School of Liberal Arts, Mercy College - Dobbs Ferry, USA
Albert M. K. Cheng, University of Houston, USA
Amar Ramdane Cherif, University of Versailles, France
Dickson Chiu, Dickson Computer Systems, Hong Kong
Gihwan Cho, Chonbuk University, Korea
Michal Choras, University of Technology and Life Sciences Bydgoszcz (UTP), Poland
Franklyn Chukwunonso, Universiti Teknologi Malaysia (UTM), Malaysia
Jose M. Claver, University of Valencia, Spain
Francesco Colace, DIEII - Università degli Studi di Salerno, Italy
Rebeca Cortázar, University of Deusto, Spain
Beata Czarnacka-Chrobot, Warsaw School of Economics, Poland
Kevin Daimi, University of Detroit Mercy, USA
Dimitrios Damopoulos, Stevens Institute of Technology, USA
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Bernard De Baets, Gent University, Belgium
José de Oliveira Guimarães, Federal University of São Carlos, Brazil
Vincenzo Deufemia, Università di Salerno - Fisciano, Italy
Alexandre Dolgui, Ecole des Mines de Saint-Etienne, France
Ludek Dolihal, Masaryk University - Brno, Czech Republic
Juan Carlos Dueñas, ETSI Telecomunicación, Universidad Politécnica de Madrid, Spain
Christof Ebert, Vector Consulting Services GmbH, Germany
Chanaka Edirisinghe, The University of Tennessee - Knoxville, USA
Hans-Dieter Ehrich, Technische Universität Braunschweig, Germany
Moez Esseghir, Technology University of Troyes, France
Fausto Fasano, University of Molise, Italy
Maria Fazio, University of Messina, Italy
Javier Dario Fernandez Ledesma, Universidad Pontificia Bolivariana - Medellín, Colombia
Kurt B. Ferreira, Sandia National Laboratories, USA
Joerg Fliege, The University of Southampton, UK
Rana Forsati, Shahid Beheshti University - Tehran, Iran
Rita Francese, Università di Salerno - Fisciano, Italy
Bernhard Freudenthaler, Software Competence Center Hagenberg GmbH, Austria
Matjaz Gams, Jozef Stefan Institute - Ljubljana, Slovenia
Raghu Ganti, IBM Thomas J. Watson Research Center, U.S.A.
Félix J. García, University of Murcia, Spain
David Garcia Rosado, University of Castilla-La Mancha, Spain
Debasis Giri, Haldia Institute of Technology, India
Apostolos Gkamas, University Ecclesiastical Academy of Vella of Ioannina, Greece
Gustavo González, Mediapro Research - Barcelona, Spain
Feliz Gouveia, University Fernando Pessoa, Portugal

Copyright Information

Table of Contents

Towards Migration of User Profiles in the SONIC Online Social Network Federation

Sebastian Göndör, Felix Beierle, Evren Küçükbayraktar, Hussam Hebbo, Senan Sharhan and Axel Küpper
Service-centric Networking
Telekom Innovation Laboratories, TU Berlin, Germany
Email: sebastian.goendoer@tu-berlin.de, beierle@tu-berlin.de, evren.kuecuekbayraktar@campus.tu-berlin.de,
hussam.hebbo@campus.tu-berlin.de, senan.mh.sharhan@campus.tu-berlin.de, axel.kuepper@tu-berlin.de

***Abstract*—As of today, even though there is a strong trend of Online Social Networks (OSNs) becoming the main communication medium, OSN platforms are still mostly proprietary, closed solutions, which are not capable to seamlessly communicate with each other. The research project SONIC (SOcial Network InterConnect) proposes a holistic standard for inter-platform communication to eradicate these gaps between OSN platforms. Yet, even with such a communication architecture, users are still bound to the platform they originally signed up with. We envision a mechanism that allows users to migrate their social profiles between OSN platforms at any time without losing any data or connections. To facilitate a seamless migration of a user's social profile from one OSN platform to another, we propose a standardized container format for social profile migration and a protocol for migration of the profile data. This allows users to move their social profiles to a new platform server without losing any data such as images or status messages. In order to uniquely identify social profiles across multiple platforms, a globally unique identifier (*Global ID*) is assigned to each profile. This way, social profiles, as well as references to such profiles can be kept intact when the location of the social profile has changed due to migration. To inform linked social profiles about the recently conducted migration of the profile, a Global Social Lookup System (GSLS) maintains a database of *Social Records*, which link the *Global ID* to the current profile location.**

***Keywords*–Online Social Networks; Social Profile Migration**

I. Introduction

Online Social Networks (OSN) have become an integral part of our everyday digital social lives. While functionality of early social platforms, such as *Classmates.com* or *Sixdegrees* was mostly limited to discussion boards and modeling and maintaining relationships to friends and colleagues, today's OSNs have become one of the main communication platforms, which allow users to communicate via text, audio, and video, share content, plan events, or just stay in contact with friends and relatives. While a large number of competing OSN platforms with a broad variety of features exist as of today, *Facebook*, which was founded in 2004, managed to overcome its predecessors and competitors by far in terms of number of users and popularity [1]. This forced many competitors to discontinue their services, or focus on niche markets, such as focusing on modeling links between business partners.

Today's OSN platforms are mostly organized in a centralized manner. This forces users of OSN platforms to not only entrust all personal information to the respective platform's operator and surrender copyrights of the profile's contents to the platform's operator, but also creates lock-in effects, so users are bound to the OSN platform they registered with. These lock-in effects are used to keep users from migrating to other OSN platforms at a later time. Personal data acquired from the users is then used e.g. for targeted advertisement, giving the users little or no control over how and what information is used [2][3][4]. Decentralized OSN alternatives such as the open source applications Diaspora *Diaspora* [5] or *Friendi.ca* [6] allow users to host their own data at an arbitrary server. Still, these approaches fail at allowing seamless communication with arbitrary other OSN platforms [7].

The research project SOcial Network InterConnect (SONIC) proposes a holistic standard for social inter-platform communication. Here, a common protocol is used to allow different kinds of OSN platforms to interact directly, while gaps between different platforms and servers, i.e., the fact that a social profile is hosted at another OSN platform, are kept hidden from users [8]. Following this approach, OSN platforms support a common API and protocol, which allows to exchange social information across platform borders, while addressing remotely hosted user accounts directly via a globally unique user identifier. The result is an *Online Social Network Federation* (OSNF), defined as a *heterogeneous network of loosely coupled OSN platforms using a common set of protocols and data formats in order to allow seamless communication between different platforms* [8]. If a user requests social content such as status updates, messages, or images from another user, the required data is retrieved from the social platform server of the targeted user and displayed directly in the user interface of the requesting user's social platform. As a result, users do not need to be aware of the fact that their friends might be using a different type of OSN. Besides the lack of a common communication protocol for OSN platforms, a standard for migrating social profiles between different OSN platforms has not yet been proposed. Such a standard would allow users to export their social profiles to another OSN platform if e.g., the terms and conditions of the former platform operator are changed. In this paper, we present ongoing research on a migration mechanism that allows users of any OSN platform in the SONIC OSNF to migrate their social profiles to another OSN platform of their choice. Connections between social profiles are kept intact and therefore allow a seamless handover.

In this paper, we present ongoing research on a migration mechanism that allows users of any compatible OSN platform in the SONIC OSNF to migrate their social profiles to another OSN platform of their choice. Connections between social profiles are kept intact and allow a seamless handover. In Chapter II, an overview of existing standards is provided. Chapter III describes details of the migration mechanism, while Section IV concludes the paper.

II. Related Work

Several existing OSN platforms offer functionality that allows a user to export and download parts of or even the complete social profile. The motivation behind this kind of export functionality is rarely to allow the social profile to be moved to a different OSN platform, but rather to enable users to create a local profile backup. For example, Facebook, Google+ and Twitter offer export functionality that saves accumulated data into an archive, which can be downloaded by the user. Other networks such as LinkedIn only offer mechanisms to export contacts as a .csv or .vcf file, yet an export mechanism for the whole profile is not provided. Federated and open source approaches such as Friendi.ca or Diaspora also offer basic data export mechanisms. Even though profile data can be exported, the functionality has been designed for a personal data backup for user profiles and lacks an import mechanism [9]. The DataPortability Workgroup that aimed to define data formats and best practices based on open formats to exchange user account data between different platforms [10][11]. In 2008, Facebook and Google joined the Workgroup, however, as the present lock-in situation demonstrates, to no avail.

For identification of user profiles, current OSNs mostly use a locally unique identifier in combination with the platform server's domain name. Here, URLs (e.g. http://osn.com/alice) or email-like identifiers (e.g. alice@osn.com as with XMPP, Diaspora, or Friendi.ca) are the most common formats. The resulting identifiers are globally unique, but bound to the platform's domain name. Hence, connections to other social profiles would be lost when migrating a profile to another server, as the user identifier needs to be changed to reflect the change in location. Universally Unique IDentifiers (UUIDs), are a general format concept to generate 128-bit globally unique identifiers in a distributed fashion. Here, hash functions are used to reduce the probability of a collision to a minimum, therefore allowing for UUID generation without a central entity or directory [12]. Similar approaches such as Twitter Snowflake or boundary flake have been proposed, which are based on the same principle. As these IDs do not comprise location information, additional directory services are required to resolve a UUID to a URL.

Even though several OSN platforms allow exporting parts of a user's social profile, functionality for uploading a previously downloaded social profile is missing, thereby rendering it impossible to migrate or restore social profiles. Furthermore, user identifiers are mostly bound to a certain domain, which would result in a loss of links to other social profiles when the location of a profile changes. To this point, a holistic mechanism that allows migration of complete social profiles has not yet been proposed.

III. Profile Migration

Implementing inter-platform communication through standardized protocols and data formats as proposed by SONIC allows users of OSN platforms to freely choose, which platform operator they want to entrust their data to. However, once set up at a certain platform, a social profile cannot be moved. Although some OSN platforms allow to export (parts of) a user's profile data, importing this data into another OSN platform is usually either impossible or cumbersome and has to be done manually. Moreover, the most severe drawback is that links to other users are lost, as a change of the OSN platform results in a different URL of the profile. In order to truly eradicate lock-in effects of today's OSN platforms and allow users to become fully independent of any OSN provider, users need to be able to migrate their social profiles between different OSN platforms in a standardized and automated fashion without losing connections to and from other social profiles. This way, users are enabled to reconsider their choice of an OSN platform and move their social profiles, if for example the OSN platform's terms of usage service are changed, or the platform operator goes out of business. The research project SONIC proposes a set of social container formats, which are designed to store social profile data including profile pages, exchanged messages, status updates, or images. By providing a common protocol for extraction of a social profile as well as importing the data on any compatible target OSN platform, social profiles can be migrated between platform servers automatically. To further allow links between different social profiles to be kept intact when migrating to other servers, SONIC proposes to assign a *Global ID* to each social profile. *Global IDs* are domain independent and globally unique identifiers that allow to address social profiles independently of their URL. Information about the current location of a social profile is specified in a *Social Record*, which links the *Global ID* to it's actual URL. Finally, all *Social Records* are published and stored in the Global Social Lookup System (GSLS). The GSLS is a decentralized and global directory service, that allows OSN platforms to resolve a *Global ID* to retrieve the current location.

A. Global Identifiers

Social profiles are highly complex datasets, which are interlinked with each other. Traditionally, social profiles are identified via a username, which is unique only for the hosting OSN platform. In combination with the OSN platform's domain name, a globally unique, but domain-dependent user id is created. Therefore, changing a social profile's location would break links to other users' social profiles as a result of the changed profile location. This problem extends beyond the friend rosters, as social profiles are commonly linked in content such as status updates, conversations, or images. Therefore, SONIC proposes the use of a *Global ID* as a domain independent and globally unique identifier, which is kept intact and unchanged during the migration process. As depicted in Figure 1, the *Global ID* is a 256-bit sequence created by using PBKDF2 [13] with 10,000 iterations using the user's account public key, and a salt of fixed length. The output is converted to base36 ([A-Z0-9]) to shorten it for the use on screen and in URLs. As *Global IDs* do not comprise information about the location of a social profile, SONIC proposes the GSLS as a global directory service, which allows to resolve a *Global ID* to it's matching *Social Record*. *Social Records* comprise information about a social profile's current location, as well a digital signature as proof of authenticity and integrity.

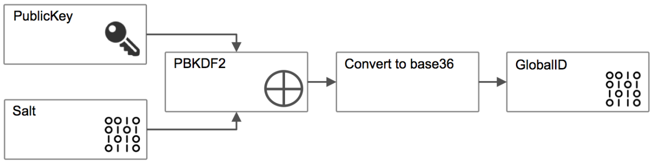


Figure 1. Deriving the GlobalID from the *Personal PublicKey* using PBKDF2

B. Social Records

Social Records are JSON encoded datasets, which contain a set of information denoting - among other data - the current social profile's location. Each Social Record dataset is uniquely identifiable by it's *Global ID*, which is used as a globally unique lookup key in the GSLS. By retrieving a *Social Record* for a known *Global ID* from the GSLS, information about the current location of the associated *Social Profile* can be retrieved. All *Social Records* are stored in a decentrally organized directory service. This directory service, the GSLS, uses a DHT to distribute the *Social Records*. When retrieving the *Social Record* for a *Global ID*, authenticity and integrity of the data can be verified by a digital signature, which is part of the *Social Record*. The public key required for verification of the signature is also included in the dataset. This way, data corruption or intentionally altered *Social Record* datasets can be detected. Each *Social Record* contains the public keys from two key pairs associated with the according social profile:

- **Personal KeyPair** The *Personal KeyPair* is used to create the *Global ID* of a *Social Account* and sign the *Social Record*. The *Personal KeyPair* cannot be revoked or exchanged, as any change of the public key would result in a new *GlobalID*.
- **Account KeyPair** The *Account KeyPair* is used and managed by the *Platform* to sign content created by the account owner as well as requests and responses. As this requires that the private part of the *Account KeyPair* is entrusted to the platform, the *Account KeyPair* can be revoked and exchanged with a new key pair. When exchanging a key pair, revocation information is published as part of the *Social Record*, which is signed using the *Personal KeyPair*.

Exchanging the cryptographic keys of the *Social Record* would allow an attacker to alter the included data and create a valid digital signature. To prevent this attack scenario, the *Global ID* is derived directly from the public key and a salt of fixed length. Therefore, exchanging the cryptographic keys of the *Social Record* would automatically alter the *Global ID*. As the *Global ID* is used as the lookup key in the directory service, an exchange of the *Personal KeyPair* is rendered impossible. In order to allow revocation of the *Account KeyPair*, the *Personal Keypair* is used to both create the *Global ID* and sign the *Social Record*. This way, the *Global ID* remains unchanged when the *Account KeyPair* needs to be revoked. Revocation information is also stored in the *Social Record* and signed using the *Personal KeyPair*. The *Social Record* comprises the following information:

- **Global ID** The *Global ID* is the identifier of the *Social Record* as well as the social profile. It is a 256-bit sequence created from the public key of the *Personal KeyPair* and a salt.
- **Salt** A sequence of 16 random characters.
- **Account PublicKey** Digital signatures for content, requests, and responses are created using the *Account KeyPair*. To allow other users to verify these signatures, the *Social Record* contains the public key of the *Account KeyPair*.
- **Personal PublicKey** The *Personal PublicKey* is used to create the *Global ID* of a social profile and sign the contents of the *Social Record*. Additionally, the *Personal KeyPair* is used for key revocation. As the signature of the *Social Record* is created using the private key of the *Personal KeyPair*, the *Personal PublicKey* can be used to verify the integrity and authenticity of the *Social Record*.
- **Profile Location** Specifies the URL of the SONIC API endpoint at which the profile is reachable. Using the SONIC protocol, the social profile and associated resources such as images and status updates can be requested via the *Profile Location*.
- **Timestamp** Denotes the date and time of the last change to the *Social Record*. Specified in XSD-DateTime format.
- **Display Name** Human-readable name of the owner of the social profile, which is used for on-screen display.
- **Key Revocation List** List of (potentially) compromised *Account PublicKeys*. Similar to the X.509 CRL standard [14], the list describes the revoked public key, date of revocation, and a code denoting the reason of revocation.
- **Signature** A digital signature created by using the *Personal KeyPair*. The signature covers the entire *Social Record*, rendering forging of the data impossible.
- **Active Flag** Specifies, whether the *Social Record* is active or not. A complete deletion of a *Social Records* is not possible to prevent reissuing of *Global IDs*. This way, creation of *Social Records* for already existing, yet inactive *Global IDs* is prevented.

C. The Global Social Lookup System

In order to publish *Social Records* for all social profiles, SONIC proposes the GSLS as a directory service. As SONIC proposes an open and decentralized architecture, the directory service is organized as a distributed database inspired by a DHT-based DNS alternative by Ramasubramanian and Sirer [15]. This approach provides a similar performance as the traditional hierarchical DNS, but is far more resilient against attacks [16]. The implementation of the GSLS is based on Java, where a Jetty server provides a RESTful interface for requests. Internally, TomP2P [17] is used to build and maintain the DHT, which is responsible for storage and replication of the *Social Records*. The external API features support for retrieving, creating, updating, and deleting *Social Records*:

- **READ** Retrieves a *Social Record* for a *Global ID* specified in the request.
- **CREATE** Allocates a previously not occupied *GlobalID* for a new *Social Record*. A correctly formatted and signed updated version of the *Social Record* for this *Global ID* has to be provided.
- **UPDATE** Updates an existing *Social Record* in the GSLS. A correctly formatted and signed updated version of the *Social Record* for this *Global ID* has to be provided.
- **DELETE** Disables a *Social Record* for a given *GID*. The *Social Record* is just deactivated in order to prevent the *Global ID* from being claimed by other users.

This interface allows not only to retrieve information about a social profile's current location, but also supports migration of social profiles. In case a profile is migrated to a new OSN platform, the *Social Record* can be updated accordingly. Requests to this profile are then automatically redirected to its new location.

D. Generic Data Formats

In order to support cross-platform exchange of social profiles and associated information, SONIC proposes generic data formats for social profiles in JSON, which have been extended and adopted for the purpose of migration. In SONIC, a standardized set of core features has been identified that covers functionality provided by the broad majority of all OSN platforms. This feature set covers all basic functionality of OSN platforms being profile pages, friend rosters, status messages, liking content, commenting on content, tagging people, and multi-user conversations, but can be extended easily to support additional functionality [8]. These data formats were designed based on previous analysis, comparison, and mapping of different OSNs' features in order to ensure the greatest possible coverage of standard OSN features.To facilitate profile migration, SONIC proposes a suitable migration data structure that holds the data that is common to most existing OSN platforms, which is based on the regular data formats of SONIC but has been extended and adopted for the purpose of migration. Here, every JSON-encoded message type corresponds to a table of this proposed data structure to ease the extraction of data from an OSN platform. To validate incoming and outgoing data, every JSON-encoded message type has a corresponding JSON schema type [18] to ensure conformity with the format as a security measure to avoid accepting malformed data.

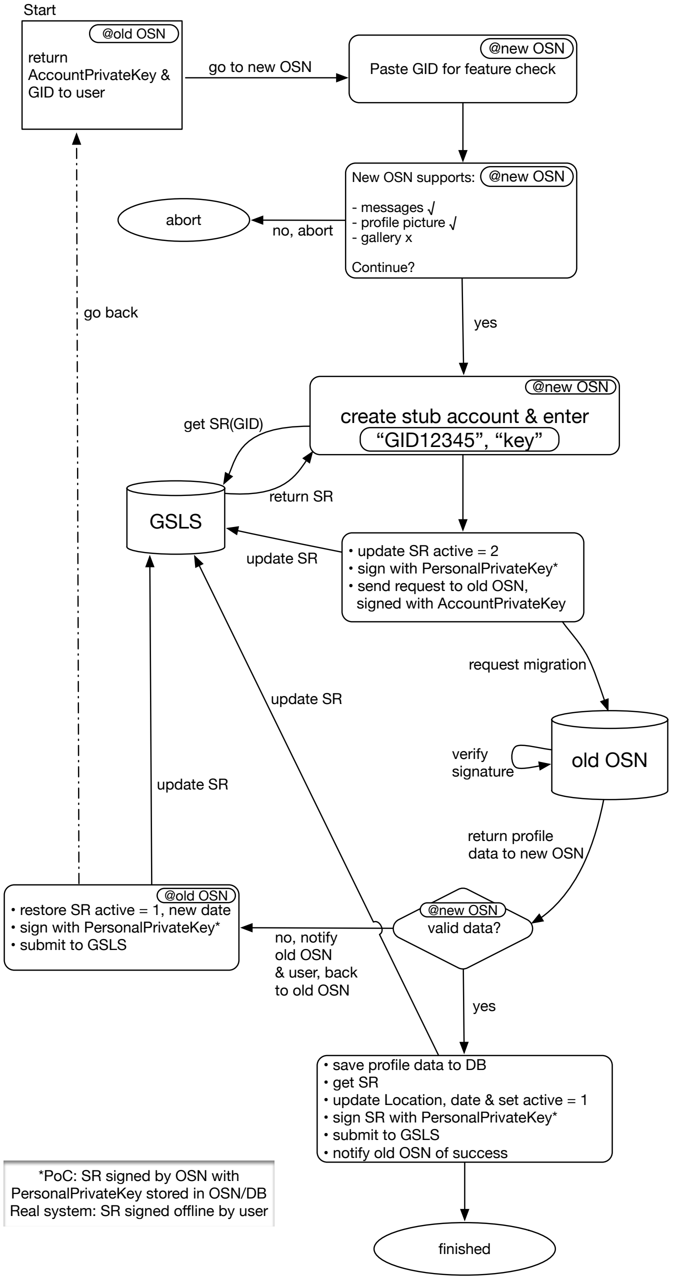

Start
@old OSN
return AccountPrivateKey & GID to user
go to new OSN
@new OSN
Paste GID for feature check
New OSN supports:
@new OSN
- messages √
- profile picture √
- gallery x
Continue?
abort
no, abort
go back
yes
@new OSN
create stub account & enter
"GID12345", "key"
get SR(GID)
return SR
GSLS
update SR
• update SR active = 2
• sign with PersonalPrivateKey*
• send request to old OSN, signed with AccountPrivateKey
request migration
update SR
verify signature
old OSN
update SR
return profile data to new OSN
@old OSN
• restore SR active = 1, new date
• sign with PersonalPrivateKey*
• submit to GSLS
no, notify old OSN & user, back to old OSN
@new OSN
valid data?
yes
• save profile data to DB
• get SR
• update Location, date & set active = 1
• sign SR with PersonalPrivateKey*
• submit to GSLS
• notify old OSN of success
*PoC: SR signed by OSN with PersonalPrivateKey stored in OSN/DB
Real system: SR signed offline by user
finished

Figure 2. Migration flow-chart: After providing both Global ID and the key pair to the new platform, all profile data is fetched from the old OSN platform using standard data formats. Finally, the profile location is updated.

E. Migration Protocol

The proposed migration protocol, as depicted in Figure 2, describes the necessary steps to migrate a social profile to a new platform. Before migrating a social profile, the user has to select a new OSN platform as a migration endpoint to which the social profile will be migrated to. At the new OSN Platform, he provides the information necessary for authentication at the old OSN platform, at which the social profile is stored. This information comprises the *Global ID* and the *Account PrivateKey*. Using the provided *Global ID*, the new OSN can now retrieve the *Social Record* of the user. Now the old OSN platform is queried for a list of supported features using `GET /features`, indicating which profile data is incompatible with the new OSN platform and therefore cannot be used or displayed after a successful migration. Only if the user still consents and confirms the migration to the new OSN platform proceeds. Otherwise, the operation is aborted. If the user agrees to proceed, the provided information is used to create a stub profile at the new OSN platform. All social profile data will be migrated to this stub profile. The user now updates the *Social Record* by setting the active flag to `active = 2` to indicate that the profile is currently being migrated and requests from other OSN platforms cannot be handled at the moment. The new OSN platform now requests the migration data from the old OSN platform using a standard remote procedure call `HTTP GET /:globalID/migration`. The request is signed with the *Account PrivateKey*, so the request's authenticity can be verified by the old OSN platform. The old OSN responds with the JSON-encoded profile data, which is validated against the specified JSON schemas and, on success, saved to the new OSNs database. Once all social profile data has been written to the new OSN platform's database, the migration is concluded by sending a notification to the old OSN platform using `HTTP PUT /:globalID/migration`, which then deletes all data. The *Social Record* is once more updated by setting the active flag to `active = 1` and updating the profile location to reflect the profile's new URL, thus concluding the migration process. If, at any moment, the migration is aborted, the new OSN platform will send a failure notification to the old OSN platform. In this case, data at the old platform will not be deleted and the *Social Record* in the GSLS will be updated by setting the active flag back to `active = 1`. Ultimately, the new OSN platform will delete all received profile data and provide a detailed error description in order to allow tracing the cause of the migration failure. To prevent misuse of the *Account KeyPair* by the old OSN platform, the key pair is revoked as an optional security measure. Upon key revocation, the new *Account KeyPair* is created by the new OSN platform, while the new *Account PublicKey* as well as the revocation certificate for the old key are published in the GSLS.

IV. Conclusion

In this paper, we described the current state of ongoing work on a migration protocol for social profiles in the SONIC OSNF. The migration protocol proposes a set of generic data formats based on widely accepted web standards to encapsulate all information related to a social profile. Using common APIs for export and import, social profiles can be migrated between compatible OSN platforms in an automated fashion. As the architecture of SONIC proposes the use of globally unique and domain-independent identifiers (*Global IDs*) for social profiles, connections to and from social profiles will not break when migrating. Profile locations are published in *Social Records*, which link a profile's *Global ID* to its current location. Profile lookup is facilitated through the Global Social Lookup System (GSLS), a DHT-based directory service. This way, connections between social profiles are kept intact even when the location of a profile is changed. The solution is currently being implemented and tested in a SONIC testbed running a specially enhanced version of *Friendi.ca* [6], which supports the SONIC protocol.

Acknowledgment

The work described in this paper is based on results of ongoing research as part of the research project SONIC (grant number `01IS12056`). The project's web page is accessible at http://sonic-project.net. *SONIC* is funded as part of the *SoftwareCampus* initiative by the *German Federal Ministry of Education and Research (BMBF)* in cooperation with *EIT ICT Labs Germany GmbH*.

References

[1] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The Anatomy of the Facebook Social Graph," arXiv preprint arXiv:1111.4503, 2011.

[2] A. Datta, S. Buchegger, L.-H. Vu, T. Strufe, and K. Rzadca, "Decentralized Online Social Networks," in Handbook of Social Network Technologies and Applications. Springer, 2010, pp. 349–378.

[3] B. Fitzpatrick and D. Recordon, "Thoughts on the Social Graph," 2007, http://bradfitz.com/social-graph-problem/ [accessed: 2015-08].

[4] W3B, "Trends im Nutzerverhalten," 2013, http://www.fittkaumaass.de/reports-und-studien/trends/nutzerverhalten-trends [accessed: 2015-08].

[5] Diaspora, "Diaspora Website," 2015, http://joindiaspora.com [accessed: 2015-08].

[6] Friendi.ca, "Friendi.ca Website," 2015, http://friendica.com/ [accessed: 2015-08].

[7] A. Bleicher, "The Anti-Facebook," IEEE Spectrum, vol. 48, no. 6, 2011, pp. 54–82.

[8] S. Göndör and H. Hebbo, "SONIC: Towards Seamless Interaction in Heterogeneous Distributed OSN Ecosystems," in Wireless and Mobile Computing, Networking and Communications (WiMob), 2014 IEEE 10th International Conference on. IEEE, 2014, pp. 407–412.

[9] Diaspora, "Diaspora FAQ," 2015, https://wiki.diasporafoundation.org/FAQ_for_users [accessed: 2015-08].

[10] K. Heyman, "The move to make social data portable," Computer, vol. 41, no. 4, April 2008, pp. 13–15.

[11] DataPortability Project, "DataPortability Project Website," 2012, http://www.dataportability.org [accessed: 2015-08].

[12] P. Leach, M. Mealling, and R. Salz, "RFC4122: A Universally Unique IDentifier (UUID) URN Namespace," 2005, http://www.ietf.org/rfc/rfc4122.txt [accessed: 2015-08].

[13] B. Kaliski, "RFC 2898: PKCS# 5: Password-based Cryptography Specification Version 2.0," 2000.

[14] D. Cooper, S. Santesson, S. Farrell, S. Boeyen, R. Housley, and W. Polk, "RFC5280: Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile," 2008, http://tools.ietf.org/html/rfc5280 [accessed: 2015-08].

[15] V. Ramasubramanian and E. G. Sirer, "The Design and Implementation of a Next Generation Name Service for the Internet," ACM SIGCOMM Computer Communication Review, vol. 34, no. 4, 2004, pp. 331–342.

[16] D. Massey, "A Comparative Study of the DNS Design with DHT-Based Alternatives," in In the Proceedings of IEEE INFOCOM'06, vol. 6, 2006, pp. 1–13.

[17] T. Bocek, "TomP2P, a P2P-based high performance key-value pair storage library," 2012, http://tomp2p.net [accessed: 2015-08].

[18] F. Galiegue and K. Zyp, "JSON Schema: Core Definitions and Terminology," Internet Engineering Task Force (IETF), 2013.

Active Intrusion Management for Web Server Software: Case WordPress

Patrik Paarnio
Department of Business Management and Analytics
Arcada University of Applied Sciences
Helsinki, Finland
e-mail: patrik.paarnio@gmail.com

Sam Stenvall
Department of Business Management and Analytics
Arcada University of Applied Sciences
Helsinki, Finland
e-mail: sam.stenvall@nordsoftware.com

Magnus Westerlund
Department of Business Management and Analytics
Arcada University of Applied Sciences
Helsinki, Finland
e-mail: magnus.westerlund@arcada.fi

Göran Pulkkis
Department of Business Management and Analytics
Arcada University of Applied Sciences
Helsinki, Finland
e-mail: goran.pulkkis@arcada.fi

***Abstract*—Methods for active management of intrusion attacks against WordPress web sites are proposed for improved real-time web security. Intrusion management is defined to be active when both intrusion responses and forensic investigations are proactive and/or automatically triggered by intrusion attacks. Booby traps as active defense against intrusion attacks using return-oriented programming and other related research is briefly surveyed. Active intrusion management techniques such as booby trapped patches to publicly known vulnerabilities in WordPress plug-ins and redirection scripts for WordPress plug-ins are proposed. Experimentation results with proposed booby trapped patches and proposed redirection scripts are presented and evaluated.**

Keywords – active intrusion prevention; active intrusion detection; web site vulnerability; WordPress vulnerability; booby trap.

I. INTRODUCTION

A fully secure network must be able to resist any type of intrusion attack and all vulnerabilities in the network must be eliminated, while it is sufficient for an attacker to find only one network weakness. Current defense methods, such as firewalls, antivirus software, and intrusion detection systems (IDS) cannot prevent all types of intrusion attacks. Most current defense methods react rather passively on intrusion attacks with intrusion alert messages to a human network administrator or to a computer doing network administration. Thus, a current IDS can often only detect occurred intrusion and related network damage. Forensic information of an intrusion attack can also usually only be traced afterwards from log files and from other system state changes caused by the attack.

Securing or hardening Content Management Systems (CMS) has become a struggle for web site administrators. The exploitation of CMS systems such as Joomla and WordPress is extraordinarily easy due to rapid development of plug-ins for the systems, bad software engineering practices (e.g., lack of quality assurance for plug-ins), and the ease of use (in-depth technical skills are not required for installing and using the software). This allows a potential attacker to scan for public installations and their corresponding vulnerabilities with minimal risk to be detected as a threat. We consider current passive intrusion management methods often too limited in ability to secure installations.

Intrusion management is defined to be active when both intrusion responses and forensic investigations are proactive and/or automatically triggered by intrusion attacks. This paper presents active intrusion management of web server software based on WordPress. We present using an exploratory case study methodology for developing an understanding of the underlying system deficiencies. This method allows us to gain deeper insights into chains of cause and effect, in the specific software, to answer the research question of how to improve security in WordPress through active intrusion management.

The paper is organized as following; we start with related research by introducing return-oriented programming as a method for both performing attacks and as a defense mechanism for implementing “booby traps”. The third section then develops an analogy for open source web software based on a case study for WordPress. In the fourth section we report preliminary research results, before concluding in the final section.

II. RELATED RESEARCH

Prevention, detection, and responding to intrusion attacks has for many years been an important research topic. Intrusion responses are created through notification, manually, and automatically. Four desirable features of an ideal intrusion response system have been proposed: automated responses, proactivity, adaptability, and cost efficiency [1].

Active intrusion defense is based on automated intrusion responses. In [2], it is proposed an intrusion management system based on intelligent decision making agents invoking response executables and scripts for different intrusion attack types. Active defense based on distributing new access control policies to firewall nodes in a network once intrusion is detected is presented in [3].

Active defense called honey-patching against attempt to exploit network software vulnerabilities is proposed in [4]. A honey-patch redirects attacks to an unpatched decoy, which collects relevant attack information and also allows attacks to succeed in order to deceive attackers.

A Linux distribution based on Ubuntu LTS, the Active Harbinger Distribution, includes many preinstalled and configured tools for active defense against malicious activity such as network scanning and connecting to restricted services. The functions of these tools "range from interfering with the attackers' reconnaissance to compromising the attackers' systems. [5]

In a guide to intrusion detection and prevention systems, automated intrusion attack responses are characterized as a technique, which "can respond to a detected threat by attempting to prevent it from succeeding". Such intrusion attack responses can

- stop the intrusion attack by terminating the network connection or user session being used by the attack or by blocking all access to the target of the attack,
- change the security environment of the target of the intrusion attack, for example by reconfiguring a firewall or a router or by applying a patch to a vulnerability exploited by the attack, or
- change the intrusion attack process from malicious to benign, for example by removing malicious file attachment from e-mail messages before they reach their recipients [6].

Active defense called "booby trapping" against code-reuse intrusion attacks based on return-oriented programming (ROP) [7] is presented in [8].

A. Return-Oriented Programming (ROP)

In a ROP attack the attacker takes over program flow control in a network connected computer without injection of malicious program code. ROP gadgets i.e., short instruction sequences terminating with a RETN assembly instruction (return from procedure) are linked together from the control stack. RETN fetches the return address from the control stack, which is manipulated in a ROP attack.

A ROP attack requires some buffer overflow vulnerability. The attack starts with injection and execution of program code, which overwrites a return address of a RETN instruction on the control stack. The resulting execution of RETN is a jump to another gadget selected by the attacker. The ROP attack is implemented by execution of a gadget chain. [7][9]

A ROP attack can succeed only if two preconditions are fulfilled: the flow control of a program must be acquired and in the program there must be gadgets, which can be linked together in an attack. A ROP attack is prevented when at least one precondition is absent. Elimination of all buffer overflow possibilities in a program prevents a ROP attack to start. Replacement of all RETN instructions in a program with other subprogram return instructions prevents a ROP attack to proceed. An indirect defense against ROP attacks is Address Space Layout Randomization (ALSR), in which instruction addresses in the program are changed. Then, a ROP attack cannot find needed gadgets at expected addresses, but can still search needed gadgets with brute force methods [10]. To detect an ongoing ROP attack the RETN execution frequency can be monitored to issue an alert if a preset threshold value is exceeded [11] [12].

B. Booby Trapping

Booby trapping software is active defense based on ROP functionality against intrusion attacks using ROP. A booby trap is program code inserted into a computer program as binary gadget changes at compile time or at load time of binaries in such a way that the functionality of the program remains unchanged. A booby trap can thus be executed only by ROP attacks against the changed program. A booby trap is program code and cannot therefore be deactivated by an intrusion attack. The program code of a booby trap can send an alert to a network administrator, register the IP address of the intrusion attack source, redirect the attack to a honeypot, even launch a counterattack, etc. [8]

Insertion of booby traps at compile time requires access to the original source code of a program. Insertion of booby traps at load time is possible using suitable machine code jump instructions without access to the original source code. Booby traps can be inserted at binary addresses of exploitable gadgets or randomly to possibly trap intrusion attacks, which scan programs to find exploitable gadgets. Figure 1 illustrates binary code changes with inserted booby traps in a program [8].

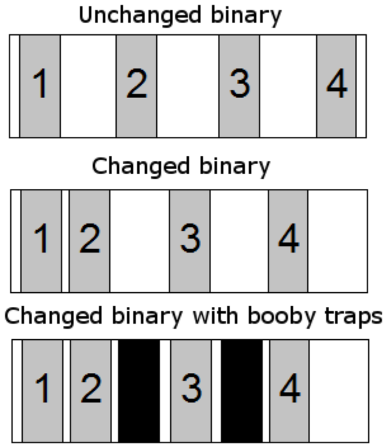


Figure 1. Binary changes with insertion of two booby traps in a program with four exploitable gadgets.

III. ACTIVE INTRUSION MANAGEMENT FOR WEB SERVERS – CASE WORDPRESS

The basic concept of ROP attacks does generally not apply to web applications. While ROP attacks against the Apache web server itself have been implemented and evaluated [13], protection against such attacks will not prevent attacks against web server code. Web server application code is easily booby trapped by modifying the source code, since the code is generally not pre-compiled.

A WordPress [14] installation on Apache or on some other web server platform has several known vulnerabilities, which have been patched. It can also be considered highly likely that there are several still publicly unknown zero-day vulnerabilities prone to intrusion attacks. The reasons of the high vulnerability of WordPress is the modular software architecture with a multitude of possible, potentially vulnerable plug-ins and the open source code, which anyone can examine to find exploitable vulnerabilities. The large global WordPress user base is also a stimulating feature for intrusion attacks against WordPress installations.

Program vulnerability patches can be booby trapped to trigger collection of forensic information about intrusion attack attempts based on available exploits. Collected forensic information can be used to create proactive responses to possible future intrusion attacks, for example by blacklisting source IP addresses related to detected use of exploits. Attempts to exploit patched or even publicly unknown vulnerabilities in non-existing (i.e., not installed) plug-ins can be redirected to honeypots or to sandbox environments where responses and/or forensic investigations are automatically triggered. An Internet connected WordPress installation with patched vulnerable plug-ins on an Apache web server has been used in booby trapping experiments described in this chapter

A. *Booby Trapped WordPress Vulnerabilities*

Vulnerabilities are patched before being booby trapped. A potential intruder shouldn't know that vulnerabilities have been patched [4]. The vulnerable version of a module is used with manual source code changes. Source code comparison of the vulnerable version with the patched version shows how the source code should be changed. A booby trap to register forensic information in a text file [15], for example logging IP addresses of intrusion attack attempts, is included in the beginning of the changed source code. The PHP code of such a booby trap is seen in Figure 2 and is denoted in later examples by booby_trap().

1) *WordPress Wp Symposium 14.11:*

The vulnerable file UploadHandler.php (see Figure 3) in the plug-in WordPress Wp Symposium 14.11 accepts for upload files of any type, which means that a malicious shell code file can be uploaded.

In [16], an exploit script is published, which creates a backdoor to protected files on a WordPress site with an uploaded shell code.

```
$ipadress = $_SERVER['REMOTE_ADDR'];
$webpage = $_SERVER['SCRIPT_NAME'];
$browser = $_SERVER['HTTP_USER_AGENT'];
$file = 'attack.log';
$fp = fopen($file, 'a');
$date = date('d/F/Y h:i:s');
fwrite($fp, $ipadress.' - ['.$date.'] '.$webpage.' '.$browser."\r\n");
```

Figure 2. Booby trap code.

```
class UploadHandler {
  …
                    'inline_file_types'=>
  '/ \.(mp4|zip|doc|docx|ppt|pptx|xls|xlsx|txt|pdf|gif|jpe?g|png)$/i',
                 'accept_file_types'=> '/.+$i',
  …
}
```

Figure 3. The vulnerable UploadHandler.php.

```
class UploadHandler {
  …
  booby_trap ( );
  'inline_file_types'=>
'/ \.(mp4|zip|doc|docx|ppt|pptx|xls|xlsx|txt|pdf|gif|jpe?g|png)$/i',
  'accept_file_types'=>
'/ \.(mp4|zip|doc|docx|ppt|pptx|xls|xlsx|txt|pdf|gif|jpe?g|png)$/i',
  …
}
```

Figure 4. The patched UploadHandler.php with an inserted booby trap.

The exploit is a Python script, which is tested before the plug-in is booby trapped. The exploit script gives the name and path of the backdoor if the upload succeeds. Using the backdoor, the file passwd can be retrieved with the command *?cmd=cat+/etc/passwd.*

Shell code upload is prevented by applying the patch shown in Figure 4. Attempts to upload files of unpermitted types create error messages after patching. A booby trap in the beginning of the patch code logs the IP addresses of attempts to exploit the patched vulnerability.

2) *WordPress Shopping Cart 3.0.4:*

The file banneruploaderscript.php in the plug-in WordPress Shopping cart 3.0.4 should check that only a logged in administrator is allowed to upload files to a WordPress site. However, any logged in user is allowed to upload files since in the condition of the if-statement is an *or-clause* instead of an *and-clause* (see Figure 5).

The published exploit script in [17] is a web form, which uses banneruploaderscript.php to upload files into the folder ./wp-content/plugins/wp-easycart/products/banners/ of a WordPress site. Upload of files of any type is allowed since banneruploaderscript.php trusts administrators and therefore an attacker is allowed to upload malicious files.

The vulnerability is patched by changing the *or-clause* to an *and-clause* in the condition of the *if-statement* in banneruploaderscript.php. After patching, a booby trap, which logs the IP addresses of attempts to exploit the patched vulnerability, is inserted (see Figure 6).

```
$userresult = mysql_query($usersqlquery);
$users = mysql_fetch_assoc($userresult);
if ($users || is_user_logged_in( )) {
  $filename = $_FILES["Filedata"]["name"];
  $filetmpname = $_FILES["Filedata"]["tmp_name"];
  $fileType = $_FILES["Filedata"]["type"];
  $fileSizeMB = ($_FILES["Filedata"]["size"] / 1024 / 1000);}
```

Figure 5. The vulnerable banneruploadscript.php.

```
$userresult = mysql_query($usersqlquery);
$users = mysql_fetch_assoc($userresult);
booby_trap( );
if ($users && is_user_logged_in( )) {
  $filename = $_FILES["Filedata"]["name"];
  $filetmpname = $_FILES["Filedata"]["tmp_name"];
  $fileType = $_FILES["Filedata"]["type"];
  $fileSizeMB = ($_FILES["Filedata"]["size"] / 1024 / 1000);}
```

Figure 6. The patched banneruploadscript.php with an inserted booby trap.

```
function mfbfw_admin_options( ){
  $settings = get_option('mfbfw');
  if ( isset($_GET['page']) && $_GET['page'] ==
    'fancybox-for-wordpress' ) {
    if ( isset($_REQUEST['action']) && 'reset' ==
      $_REQUEST['action']) {
      $settings = stripslashes_deep($_POST['mfbfw']);
      $settings = array_map('convert_chars', $settings);
      update_option( 'mfbfw', $settings );
      wp_safe_redirect( add_query_arg('reset', 'true') );
      die; }}}
```

Figure 7. The vulnerable function mfbfw_admin_options.

```
jQuery("a.fancybox").fancybox({
…
'padding': </script><script>alert(Owned by someone) </script>,
…
});
```

Figure 8. An injected piece of JavaScript.

3) *Fancybox for WordPress:*

The plug-in Fancybox for WordPress 3.0.2 has a cross-site scripting (XSS) vulnerability in the function mfbfw_admin_options (see Figure 7) in the file fancybox.php. This function doesn't validate input data and therefore permits script injection. An attacker can send the script with a web form. The script is injected on the body of a web site and is always triggered when this web page is browsed.

This vulnerability permits arbitrary JavaScript to be injected and executed on every page where Fancybox is used. The vulnerability stems from the fact that the function mfbfw_admin_options doesn't perform the necessary checks on the POST data, which allows a malicious user to inject arbitrary code into the settings for the Fancybox plug-in. The injected script is visible in the source code of the WordPress site (see Figure 8).

```
function mfbfw_admin_options( ){
  $settings = get_option('mfbfw');
  booby_trap( );
  if ( isset($_GET['page']) && $_GET['page'] ==
    'fancybox-for-wordpress' ) {
    if ( isset($_REQUEST['action']) && 'reset' ==
        $_REQUEST['action'] &&
        check_admin_referer( 'mfbfw-options-options')){
        $defaults_array = mfbfw_defaults( );
        update_option( 'mfbfw', $defaults_array );
        wp_safe_redirect( add_query_arg('reset', 'true') );
        die; }}}
```

Figure 9. The patched function mfbfw_admin_options with an inserted booby trap.

```
function wpdm_ajax_call_exec( ){
  if (isset($_POST['action']) && $_POST['action'] ==
    'wpdm_ajax_call'){
    if (function_exists($_POST['execute']))
      call_user_func($_POST['execute'], $_POST);
    else
      echo "function not defined!";
  die( );}}
```

Figure 10. The vulnerable function wpdm_ajax_call_exec.

```
function wpdm_ajax_call_exec( ){
booby_trap( );
   if (isset($_POST['action']) && $_POST['action'] ==
               'wpdm_ajax_call') {
            if ($_POST['execute']=='wpdm_getlink')
                        wpdm_getlink( );
            else
                        echo "function not defined!";
            die(); }}
```

Figure 11. The patched function wpdm_ajax_call_exec with an inserted booby trap.

In the patched source code the input data is validated by checking the source of script injection request (see Figure 9). A booby trap to log attempts to exploit the patched vulnerability is inserted before the if statement.

```
RewriteCond %{REQUEST_FILENAME} !-f
RewriteCond %{REQUEST_FILENAME} !-d
RewriteRule .* wp-boobytrap.php
```

Figure 12. Rewrite rules for redirecting requests for missing resources.

4) *WordPress Download Manager 2.7.4;*

In December 2014 a serious vulnerability in WordPress Download Manager was reported [18]. This vulnerability, for which an exploit script is published in [19] permits remote execution of program code. The script exploits the vulnerable function wpdm_cajax_call_exec (see Figure 10) in the file wpdm-core.php. The function wpdm_cajax_call_exec receives functions sent by a user from the graphical user interface and executes these functions without verification about their existence in the program. This means that an attacker can inject WordPress functions for execution in the program. The exploit script in [19] injects the WordPress function wp_Insert_user, which creates web site users for inputted user names, passwords, and user roles. An attacker can therefore create an administrator for a vulnerable WordPress site.

The functionality of the Python exploit script for WordPress DownLoad Manager 2.7.4 is tested on a vulnerable WordPress site. The WordPress site address is a parameter of the script. After successful script execution, the user information of the created administrator, which has been registered in the database of the WordPress site, is shown.

The patched source code shown in Figure 11, now permits execution of a function only if it exists in the program. A booby trap to log attempts to exploit the patched vulnerability is inserted in the patched source code. An attempt to exploit the patched vulnerability also returns an error message.

B. *Redirecting Bad Requests to a Booby Trap*

Manually booby trapping all plug-ins used on a typical WordPress installation requires much manual labor. There are simply too much possible vulnerabilities to patch, and it also makes the update process more complicated as the booby traps have to be reapplied after every plug-in update. Since booby trapping plug-ins using this approach requires the attack vector to be known (in order to insert the booby trap at the right location) it does not offer any protection against zero-day attacks.

We have studied an alternative approach that is based on two novel ideas. The first is to redirect requests for missing resources (e.g., files belonging to plug-ins that are not installed) to a special script which handles the requests. This script acts as the booby trap and can be made to do different things depending on the objective. During our testing we have configured it to log the requested URLs together with certain request parameters such as the query string and eventual POST data. If the objective would be to gather as much forensic data about potential intrusion attempts as possible the script could be programmed to emulate known exploits in order to make the attacker believe he actually succeeded. This type of emulation is often necessary since many exploits first attempt to detect whether the actual exploit would succeed or not; the payload itself may not be delivered if the detection fails.

Redirecting requests for missing files to an arbitrary request handler is not a new idea. It is used by many web frameworks, for example the Yii framework [20] and Fat-Free Framework [21], to force requests to go through the main index file of the web application itself. The same method is used here (see Figure 12), but for a different purpose.

By redirecting bad requests many types of attacks against a WordPress installation can be avoided. Even though the attacks caught by the redirection wouldn't have succeeded anyway (since the requested resource would not have been found) we have the opportunity to prevent further attacks (some of which may actually succeed) since we now can classify the IP address that made the request as malicious. This defense mechanism can thus be made a proactive part of the server's security system if it is used to automatically reconfigure the server's firewall to block further connection attempts from the implicated IP addresses.

```
# mod_substitute
AddOutputFilterByType INFLATE;SUBSTITUTE;DEFLATE
text/html text/javascript
Substitute "s|\?wpdmdl|\?fake-wpdmdl|i"
# mod_rewrite
RewriteEngine On
# stop processing when a rewrite has taken place
# and the target exists
RewriteCond %{ENV:STOP} =1
RewriteCond %{REQUEST_FILENAME} -f [OR]
RewriteCond %{REQUEST_FILENAME} –d
RewriteRule ^ - [L]
# replace the query string
RewriteCond %{QUERY_STRING} ^(.*)fake-wpdmdl(.*)
RewriteRule (.*) $1?%1wpdmdl%2 [L,E=STOP:1]
# direct requests containing wpdmdl will be
# caught here
RewriteCond %{QUERY_STRING} wpdmdl
RewriteRule .* wp-boobytrap.php
```

Figure 13. Rewrite and substitution rules for faked query strings.

Since this method of redirection is triggered only by requests for missing resources it obviously does nothing to prevent attacks against plug-ins that are installed and in use by the web site. Such attacks can potentially be mitigated using manual booby trapping, but as mentioned earlier this is very time consuming and can be error prone depending on where the booby trap has to be inserted.

We have explored the possibility of renaming installed plug-ins so that requests using a plug-in's standard URL in a potential exploit would end up being redirected to the booby trap. For this approach to be feasible, no manual modifications to the plug-ins or WordPress itself can be done, since that would make their respective update processes very cumbersome; the same modifications would have to be re-applied every time a plug-in is updated. Our research has shown that this task can be accomplished, at least partially, without editing any existing source code, using something we call faked redirection.

We have identified three ways in which an exploit may end up running code belonging to a WordPress plug-in:

- Requests directly to a file belonging to the plug-in
- Using hooks defined by the plug-in. The request goes to index.php and is internally routed to a function in the plug-in
- Using POST requests to index.php with execution paths similar to those of hooks.

The idea is to rewrite all URLs that can lead to plug-in code execution by non-standard names. Requests for the rewritten URLs would then be internally redirected to the original locations, while requests that have not been rewritten would be redirected to the booby trap.

This way, normal site usage is unaffected since all requests go through the modified URLs, but an attacker attempting to leverage an exploit against a plug-in would fail and end up in our booby trap.

The concept is easier to grasp using an example. Let's take the popular WordPress Download Manager plug-in as an example. Normally, the plug-in resides in wp-content/plugins/download-manager, and one of the hooks it uses is called wpmdl. We now substitute all occurences of wp-content/plugins/download-manager with wp-content/plugins/faked-download-manager and all occurences of wpdml with fake-wpdmdl.

Since we do not want to modify any files belonging to WordPress itself or one of the plug-ins, we use a combination of the mod_substitute and mod_rewrite Apache modules. mod_substitute is used to modify the URLs when the content is served to the browser, while mod_rewrite handles the task of reversing the substitution and eventually redirecting requests to the booby trap. Figure 13 illustrates how faked redirection is used to booby trap the wpdmdl hook that WordPress Download Manager uses.

IV. Experimental Results

Experiments with booby trapped patches to vulnerabilities in WordPress modules and with redirection scripts are presented in this chapter.

1) *Results with Booby Trapped WordPress Plug-ins:*

Intrusion attempts have produced data in log files. Some booby trapped plug-ins utilize WordPress functionality, which is reflected in the contents of log files.

The log from the booby trapped plug-in WordPress Symposium 14.11:

```
80.220.110.12 - [16/March/2015 08:17:32] /wp-
content/plugins/wp-symposium/server/php/index.php
Mozilla/5.0 (Windows NT 6.1; WOW64)
```

```
AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/36.0.1985.125 Safari/537.36
```

The log shows the path to the web page, to which the intrusion attempt has tried to upload a shell code. The exploit script fakes the information about the attacker's web browser with a predefined header. Booby trapping the patch of this vulnerable plug-in is successful, since forensic information about exploitation attempts is logged but normal administrative activities are not logged.

The log from the booby trapped plug-in WordPress Shopping Cart 3.0.4 :

```
80.220.110.12 - [01/April/2015 11:22:31] /wp-
content/plugins/wp-
easycart/inc/amfphp/administration/banneruploaderscript.php
Mozilla/5.0 (Windows NT 6.1; WOW64; rv:36.0)
Gecko/20100101 Firefox/36.0
```

The log shows the web page, which is used by the web form for file upload. The information about the intruder's web browser is public, since the intrusion attempt is made from the intruder's computer without any intermediate activity. The booby trap was not triggered by normal administrative activity.

The log from the booby trapped plug-in Fancybox for WordPress 3.0.2:

```
80.220.110.12 - [01/April/2015 10:37:07] /wp-admin/index.php
Mozilla/5.0 (Windows NT 6.1; WOW64; rv:36.0)
Gecko/20100101 Firefox/36.0
80.220.110.12 - [01/April/2015 10:49:33] /wp-
admin/plugins.php Mozilla/5.0 (Windows NT 6.1; WOW64;
rv:36.0) Gecko/20100101 Firefox/36.0
80.220.110.12 - [01/April/2015 10:49:34] /wp-admin/admin-
ajax.php Mozilla/5.0 (Windows NT 6.1; WOW64; rv:36.0)
Gecko/20100101 Firefox/36.0
80.220.110.12 - [01/April/2015 10:50:14] /wp-admin/admin-
post.php Mozilla/5.0 (Windows NT 6.1; WOW64; rv:36.0)
Gecko/20100101 Firefox/36.0
```

Fancybox is an administrative tool. All administrative activity is logged, not only the use of adminpost.php in an exploit web form. Booby trapping the patched vulnerability in Fancybox is therefore not recommended.

Part of the log from the plug-in WordPress Download Manager 2.7.4:

```
80.220.110.12 - [31/March/2015 02:58:16] /index.php
Mozilla/5.0 (Windows NT 6.1; WOW64; rv:36.0)
Gecko/20100101 Firefox/36.0
1.171.73.177 - [31/March/2015 03:26:37] /index.php
128.61.240.66 - [31/March/2015 03:53:49] /index.php
netscan.gtisc.gatech.edu
```

Each access to index.php has been logged, which includes all visits to the main web page of the WordPress site. Booby trapping a patch on the main page of a web site is not recommended.

2) Results with Redirection:

A honeypot WordPress installation was left running in an attempt to log potential exploit attempts. Redirection of request for missing resources was highly successful and many malicious requests were logged, including many aimed at exploiting software other than WordPress. Here's an excerpt showing attempts to detect vulnerable versions of two WordPress plug-ins, which were not installed on the server:

```
2015-04-03T11:45:10+00:00: Unhandled request for "//wp-
content/plugins/revslider/temp/update_extract/revslider/info.php
": $_GET = [ ], $_POST = [ ], $FILES = [ ]
2015-04-03T11:45:10+00:00: Unhandled request for "//wp-
content/uploads/wpallimport/uploads/d0bc023bca54df2d0c54efe
7b9e29311/info.php": $_GET = [], $_POST = [ ], $FILES = [ ]
2015-04-05T17:41:36+00:00: Unhandled request for "//wp-
content/plugins/revslider/temp/update_extract/revslider/info.php
": $_GET = [ ], $_POST = [ ], $FILES = [ ]
```

Initial testing shows that the faked redirection technique has the potential to catch malicious requests:

```
2015-05-20T08:36:10+00:00:
ExploitMocker\RequestHandler\Base\DefaultHandler -
Unhandled request for "/?wpdmdl=13": $_GET =
{"wpdmdl":"13"},
$_POST = [ ], $FILES = [ ]
2015-05-20T08:36:19+00:00:
ExploitMocker\RequestHandler\Base\DefaultHandler -
Unhandled request for "/?wpdmdl=15": $_GET =
{"wpdmdl":"15"},
$_POST = [ ], $FILES = [ ]
```

Without the faked redirection technique in place, the above requests would have succeeded since they are perfectly valid.

V. CONCLUSIONS

Booby trapping, originally proposed for active defense in network hosts against intrusion attacks based on return-oriented programming, has been shown to provide active forensics and proactive intrusion defense for attempts to exploit some patched vulnerabilities in WordPress web sites. However, for some vulnerabilities, booby trapping methodology must be further developed to distinguish between intrusion attempts and normal administrative activity.

If used as an active defense mechanism, redirecting requests for missing resources has the potential to catch attackers before they are able to attempt a successful exploit, assuming that the attackers have to try many exploits until they would find one that has the potential to work.

The technique is easy to implement on existing installations since it is not tied to any particular WordPress plug-ins that are installed on the server. Faked redirection can improve this even further since it can force attackers into the booby trap even if they attempt to leverage exploits that are currently unpatched. However, unlike when merely redirecting missing resources, this technique requires some manual configuration for each plug-in that the operator wants to protect.

Security in WordPress could be improved by enforcing and maintaining an internal registry for plug-ins storing their access methods, e.g. file system location. This would allow the web site administrator to randomly re-locate any installed plug-in, in a similar fashion to our faked re-direction method, without worrying that something may go wrong with the installation. This technique would essentially mitigate the effects of many zero-day vulnerabilities for WordPress installations utilizing third-party plug-ins, by allowing completely unique installation environments.

The plug-in renaming technique could be easier to implement, if plug-in authors would design their plug-ins with the possibility of renaming them in mind. The plug-ins we tested were quite unsuitable for this since their source code contained references to hard-coded plug-in URLs. The application to other similar CMS software needs to be further investigated to draw general conclusions. Still, based on our previous experience the technique holds promise as a more general solution for implementing active intrusion management into CMS software.

References

[1] N. Stakhanova, S. Basu, and J. Wong, "A Taxonomy of Intrusion Response Systems," Int. J. Information and Computer Security, vol. 1, no. 1/2, 2007, pp. 169-184.

[2] C. Carver, J. M. Hill, J. R. Surdu, and U. W. Pooch, "A Methodology for Using Intelligent Agents to provide Automated Intrusion Response," Proc. IEEE System, Man, and Cybernetics Information Assurance and Security Workshop, West Point, IEEE Press, 2000, pp. 110-116.

[3] S. Dai and Y. Du, "Design and Implementation of Dynamic Web Security and Defense Mechanism based on NDIS Intermediate Driver," Proc. 2009 Asia-Pacific Conference on Information Processing, IEEE Press, 2009, pp. 506-509.

[4] F. Araujo, K. Hamlen, S. Biedermann, and S. Katzenbeisser, "From Patches to Honey-Patches: Lightweight Attacker Misdirection, Deception, and Disinformation," Proc. 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14), ACM, 2014, pp. 942-953.

[5] ADHD provides tools for active defense, http://sourceforge.net/projects/adhd/ [retrived: August, 2015]

[6] K. Scarfone and P. Mell, "Guide to Intrusion Detection and Prevention Systems (IDPS)," Special Publication 800-94 Rev. 1 (Draft), NIST National Institute of Standards and Technology, U.S. Department of Commerce, 2012.

[7] M. Prandini and M. Ramilli, "Return-Oriented Programming," IEEE Security & Privacy, vol. 10, no 6, Nov.-Dec. 2012, pp. 84-87.

[8] S. Crane, P. Larsen, S. Brunthaler, and M. Franz, "Booby Trapping Software," Proc. New Security Paradigms Workshop (NSPW'13), ACM, 2013, pp. 95-106.

[9] R. Roemer, E. Buchanan, H. Shacham, and S. Savage, "Return-Oriented Programming: Systems, Languages, and Applications," ACM Trans. Information and System Security (TISSEC) – Special Issue on Computer and Communications Security, vol. 15, 2011, pp. 2-34.

[10] K. Onarlioglu, L. Bilge, A. Lanzi, D. Balzarotti, and E. Kirda, "G-Free: Defeating Return-Oriented Programming through Gadget-less Binaries," Proc. 26th Annual Computer Security Applications Conference (ACSAC '10), ACM, 2010, pp. 49-58.

[11] L. Davi, A.-R. Sadeghi, and M. Winandy, "ROPdefender: A Detection Tool to Defend Against Return-Oriented Programming Attacks," Proc. 6th ACM Symposium on Information, Computer and Communications Security (ASIACCS '11), ACM, 2011, pp. 40-51.

[12] R. Skowyra, K. Casteel, H. Okhravi, N. Zeldovich, and W. Streilein, "Systematic Analysis of Defenses against Return-Oriented Programming," in Research in Attacks, Intrusions, and Defenses, Lecture Notes in Computer Science, vol. 8145, Springer, 2013, pp 82-102.

[13] L. Liu, J. Han, D. Gao, J. Jing, and D. Zha, "Launching Return-Oriented Programming Attacks against Randomized Relocatable Executables," Proc. 10th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), IEEE Press, 2011, pp. 37-44.

[14] WordPress Portal. https://wordpress.org/ [retrived: August, 2015]

[15] PHPBook. How to log ip adresses in PHP, http://www.phpbook.net/how-to-log-ip-adresses-in-php.html [retrived: August, 2015]

[16] C. Viviani, WordPress Wp Symposium 14.11 - Unauthenticated Shell Upload Exploit, http://www.exploit-db.com/exploits/35543/ [retrived: August, 2015]

[17] K. Szurek, WordPress Shopping Cart 3.0.4 - Unrestricted File Upload, http://www.exploit-db.com/exploits/35730/ [retrived: August, 2015]

[18] M. Nadeau, Security Advisory – High Severity– WordPress Download Manager, http://blog.sucuri.net/2014/12/security-advisory-high-severity-WordPress-download-manager.html [retrived: August, 2015]

[19] C. Viviani, WordPress Download Manager 2.7.4 - Remote Code Execution Vulnerability, http://www.exploit-db.com/exploits/35533/ [retrived: August, 2015]

[20] Yii framework The Fast, Secure and Professional PHP Framework. http://www.yiiframework.com/ [retrived: August, 2015]

[21] Fat-Free Framework A powerful yet easy-to-use PHP micro-framework designed to help you build dynamic and robust web applications - fast! http://fatfreeframework.com [retrived: August, 2015]

Idea for Location and Detection of Fiber Cuts in PON

Mary Luz Mouronte López
Telematic and Engineering Department.
High Technical School of Engineering and Telecommunication Services. Polytechnic University of Madrid.
Spain
email: mouronte.lopez@upm.es

***Abstract*—In this paper, we present an innovative idea to build a software system that allows detecting and locating fiber cuts in Passive Optical Networks (PON) prior to the incident being reported by the customer to the provider' helpdesk. This early detection helps reduce the service restoration time and the number of customer service repair claims.**

Keywords- PON; OTDR; Fiber cut.

I. Introduction

The idea presented in this paper aims at detecting fiber cuts in advance, accelerating failure resolution, minimizing impacts on the service provision and reducing the number of claims from end customers in a PON [1][2]. We propose to design a software system that generates location information automatically in case a fiber cut occurs. At present, if this happens, users contact the provider's helpdesk, the technicians test the PON in order to find the failure point, and solve the problem on-site. The novelty of this research lies in expecting to detect the fiber cuts before the user's claim occurs, which allows reducing the elapsed time until the problem is solved and improving the customers' perception.

There are works present in the literature that address techniques used to monitor PONs. In [3], Hasegawa et al. propose pulse-Optical Coherence Domain Reflectometry (pulse-OCDR) as a new monitoring method for PONs. The work presents the results of a laboratory demonstration, where reflections from the Fiber Bragg Gratting (FBG) filters installed at the end of drop fibers in a 32-branch PON with a 15 km feeder line are identified with 2.6 cm spatial resolution, and a fault in drop fibers is detected as the absence of one FBG reflection. Although the location of a fault is not implemented because the used OCDR does not detect Rayleigh scattering [4] from the drop fibers, fault detection alone is valuable because it can tell whether the fault is in fiber or in the Optical Network Terminal (ONT) and hence reduces the number of times service personnel is dispatched on troubleshooting. In [4], Costa et al. propose a strategy to monitor the fiber plant failures by implementing an Optical Time Domain Reflectometer (OTDR) subsystem in the physical layer. The procedure is evaluated analyzing its theoretical performance limits.

Our technical solution suggests storing the signal that is reflected back from several points along the fiber in the OTDR to check later against a blueprint when a fiber cut arises. The analysis will allow to detect the cut location and to identify those customers who are impacted by the network failure.

The rest of the paper is organized as follows: Section 2 characterizes the Passive Optical Networks (PON), in Section 3 the Optical Time Domain Reflectometer (OTDR) is described, in Section 4 the method of analysis is presented, and finally in Section 5 we end with some conclusions.

II. PASSIVE OPTICAL NETWORK

A PON [1][2] is a telecommunications network that employs point-to-multipoint fiber to the premises in which unpowered optical splitters are employed to make possible a single optical fiber to serve multiple premises. A PON includes an Optical Line Terminal (OLT) at the service provider's central office and a number of Optical Network Units (ONU) close to end customers. A PON minimize the amount of fiber and central office equipment required compared with point-to-point architectures. PON technologies: Broadband Optical System Based on a Passive Optical Network (BPON) [6][7][8], Ethernet-Passive-Optical-Network (EPON) [9] and Gigabit Passive Optical Network (GPON) [10][11][12][13] are compared in Table I.

TABLE I. PON TECHNOLOGIES

Characteristics	BPON	EPON	GPON
Standard	ITU-T G.983.x	IEEE 802.3ah	ITU-T G.984.x
Typical Split ratio	32	32	64
Typical Max Reach (distance between OLT and ONU)	20 km	10 km	10-20 km
OAM	PLOAM	Ethernet OAM	PLOAM

OAM: Operation, Administration and Maintenance.
PLOAM: Physical Layer Operation, Administration and Maintenance.

III. Optical Time Domain Reflectometer

An OTDR is used to test PON, combining a laser source and a high resolution detector to give an inside view of the optical link. This device is connected to the end of the link, the laser source injects a signal into the fiber and the detector receives the light reflected from the different connection elements. The main OTDR features are: working wavelength, pulse duration, maximum measured distance, resolution (attenuation or distance), difference between the largest and the smallest value of pulse that can be measured, linearity, i.e., error resulting from nonlinearity ranges over the measured attenuation.

IV. ALGORITHM

Information about the signal that is reflected back from several points along the fiber wire can be stored by the OTDR to check later against a blueprint when a fiber cut arises. The inspection will allow to detect the cut location and to identify those users who are impacted by the network trouble. A scan should be performed over all network branches that come from the provider's central office in order to detect the failures before users perceive them. We propose a software system (test bed), which includes the following elements:

- A Personal Computer (PC).
- MatLab Tool.
- Algorithm implemented in MatLab Language.
- An OTDR, with functionality to dump data in a format file, which should be understood by MatLab.

An overview of the operation and the procedures carried out by the algorithm are shown in the following sections.

A. *Main Tasks*

Figure 1 summarizes the main actions that are taken by the algorithm.

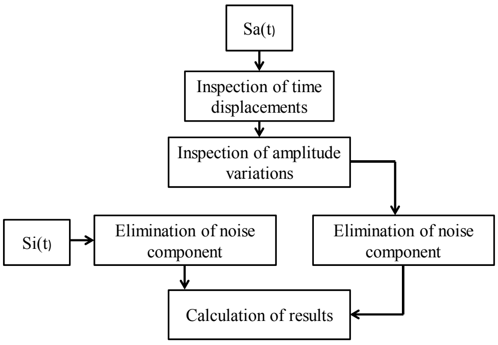

Sa(t)
Inspection of time displacements
Inspection of amplitude variations
Si(t)
Elimination of noise component
Elimination of noise component
Calculation of results

Figure 1. Main Flow Chart.

B. *Inspection of time displacements*

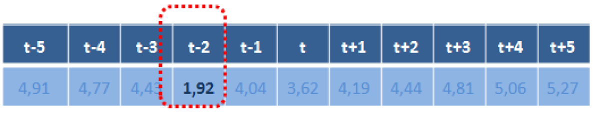
This procedure detects fluctuations of the signal values collected by the OTDR due to errors in the capture (specifically, time displacement errors). The input signal to this algorithm, Sa (t), is moved in a range D, –x/2≤D≤x/2, where x is the window size chosen by the user. Several magnitudes are calculated:

$$Sr(t)=Si(t)–Sa(t+D), D \text{ in } [-x/2, x/2] \quad (1)$$

$$\text{Average}(Sr(t)) \quad (2)$$

The output will be that signal Sfa(t)= Sa(t+D) with the lowest Average (Sr(t)).

Figure 2, shows the obtained averages for one imaginary Sr(t) with x=10.



t-5	t-4	t-3	t-2	t-1	t	t+1	t+2	t+3	t+4	t+5
4,91	4,77	4,43	**1,92**	4,04	3,62	4,19	4,44	4,81	5,06	5,27

Figure 2. Obtained averages for one imaginary Sr(t) with x=10 .

C. *Inspection of amplitude variations*

This process identifies fluctuations of the signal values collected by the OTDR due to errors in the capture (specifically, amplitude variation errors).

The input signal to this procedure, Sa(t), is moved in a range AR, –A/2≤AR≤A/2, where A is the window size chosen by the user.

$$Sr(t)=Si(t)–(Sa(t)+AR), AR \text{ in } [-A/2, A/2]$$

$$\text{Average}(Sr(t)) \quad (3)$$

The output will be that signal, Sfa(t)=Sa(t)+ AR, with the lowest Average (Sr(t)).

D. *Elimination of noise component*

Due to the noise generated in the optical fiber, the signal injected by the OTDR laser source can change significantly until that it is finally captured by the OTDR detector. This procedure eliminates this noise component.

1. In the first step, the slope of the input signal, S(t), is calculated. The following parameters are defined:
   - NS: Number of samples of the S(t), which contain all useful information.
   - W : Window size, which is selected by the user.
   - NW = Total windows (of size W) that are required to include all samples.
   - A_{HMP} = Sfa (t_{HMP}), t_{HMP} is the amplitude middle point includes in the highest window (over the total number, NW).
   - A_{LMP} = Sfa (t_{LMP}), t_{LMP} is the amplitude middle point in the lowest window (over the total number, NW).
   - $M = (A_{HMP} – A_{LMP})/NS$
2. In the second step, Y(t) = M*t and RS (t) = 2 * (S(t) – Y(t)) are calculated. Figure 3 shows the RS (t) corresponding to one imaginary signal S (t).
3. In the third step, several filters are applied, which allow displaying even clearly the peaks in RS(t).

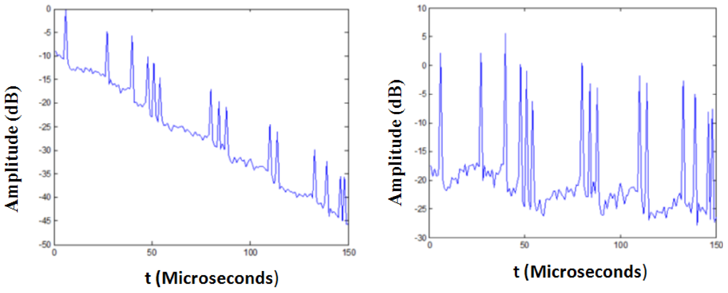


Figure 3. (Left) Imaginary signal S(t). (Right) RS(t) corresponding to S(t).

- Filter 1:

$$\text{If } RS(t) < A_{HMP} \rightarrow RS(t) = 2 * A_{HMP} \quad (4)$$

- Filter 2:

$$RS(t) = RS(t) + \text{Absolute value of } (\text{Mininum}(RS(t))) \quad (5)$$

- Filter 3:

$$\text{If } RS(t) < RS(t\text{-}1)/2 \rightarrow RS(t) = 0; \quad (6)$$
$$\text{Else } RS(t) = RS(t) \quad (7)$$

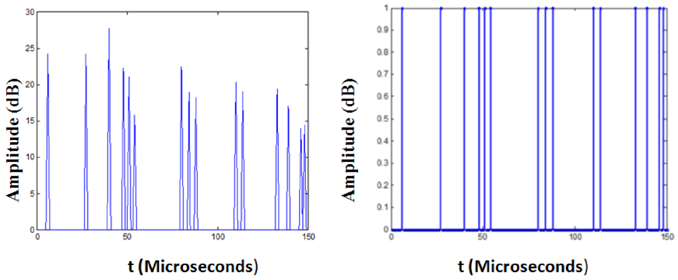


Figure 4. (Left) Imaginary RS(t) after applying the filters 1, 2 and 3. (Right) Imaginary RS(t)•

E. Calculation of results

1. In the first step, this procedure identifies the reflection occurrence, calculating :

$$\text{If } RS(t) > 0 \rightarrow RS(t)\bullet = 1 \quad (8)$$

Figure 4 (Right) shows RS (t)• for one signal imaginary RS(t).

$$C_{SfaSi\,(t)} = RS_{Sfa}(t)\bullet - RS_{Si}(t)\bullet \quad (9)$$

Δt where $C_{SfaSi\,(t)} < 0$, t corresponds to the place where a cut happens. Δt where CSfaSi (t) > 0, t refers to the location where the signal is not received due to a cut. Figure 5 shows one imaginary $C_{SfaSi(t)}$ with two peaks (one positive (t= 88) and other negative (t=66)).

2. In the second step, this procedure calculates to the provider's central office until the place where the cut occurs, that is:

Distance(Km)=(((Speed of light)*t (Δt where C_{SfaSi} (t)<0)*10^{-6})/2*RI)/1000 (10)

RI: Refraction Index
Speed of light: $3*10^{8}$ m/second

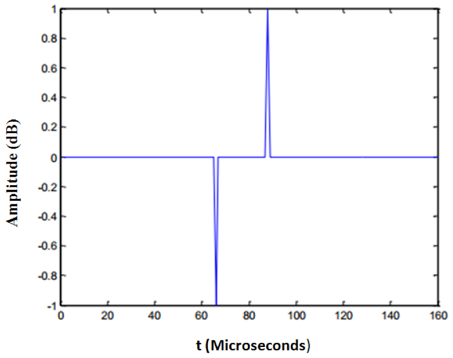
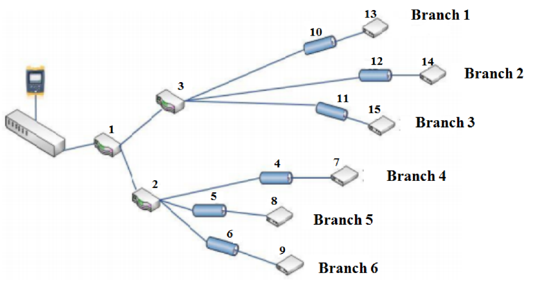


Figure 5. Imaginary CSfaSi (t).

3. In the third step, the specific branch impacted by the cut is identified. The algorithm assumes a PON composes of L levels and B braches. L and B are data provide by the user.
   a. Those values ti where $RS_{Si(ti)} = 1$ are stored in a vector V = {vj}; vj=ti ; j € N. So, for the imaginary signal displayed in Figure 5, V ={6, 27, 40, 48, 51, 54, 80, 84, 88, 110, 114, 133, 139, 146, 148}.
   b. Each value ti where $C_{SfaSi\,(ti)} > 0$ is searched in V, and its position j is deduced. So, for the imaginary signal depicted in the Figure 5 the value 88 is in j= 9.
   c. A matrix A $\{a_{lk}\}$ l<=L; k<=B, where each a_{lk} represents the element position in the PON, is built. The row and column denote the level and the branch where the device is located, respectively.

Figure 6 shows an imaginary network (in upper part) and its matrix A (in the lower part).



	Branch 1	Branch 2	Branch 3	Branch 4	Branch 5	Branch 6
Level 1	1	1	1	1	1	1
Level 2	3	3	3	2	2	2
Level 3	10	12	11	4	5	6
Level 4	13	14	15	7	8	9

Figure 6. (Upper) Imaginary network (Lower) Matrix A corresponding to the network on the upper.

d. Those a_{lk} that are equal to the positions calculated in step a, identify the not-reachable devices. Therefore, the cut must have happened in some preceding section in the PON. So, in the example shown in Figure 7, the not-reachable device is in the level 4 and the branch 6, given that $a_{46} = 9$.

V. CONCLUSION AND FUTURE WORKS

We have proposed a software system which executes an algorithm to detect and to provide information on the fiber cuts in a PON. Our future works will investigate deeper the method effectiveness (practicality per Signal Noise Rate (SNR)) testing several real scenarios.

REFERENCES

[1] H. G. Perros, "Connection-Oriented Networks: SONET/SDH, ATM, MPLS and Optical Networks", Wiley, October 2005.

[2] F. Effenberger, G. Kramer, and T. Pfeiffer, "An Introduction to PON Technologies", Topics in Optical Communications, IEEE Communications Magazine, pp. S17-S25, March 2007.

[3] T. Hasegawa and A. Inou, "Monitoring of Drop Optical Fibers in 32-Branched PON using 1.65 µm Pulse-OCDR" Sei Technical Review, vol. 69, pp. 83-85, October 2009.

[4] R. B Miles, W. R Lempert and J. N. Forkey, "Laser Rayleigh scattering", Measurement Science and Technology, February 2001, vol.. 12, pp. R33–R51.

[5] L. Costa, J. A. Lázaro, V. Pólo and A. Teixeira, "Viability of InService, Low-Cost and Spatially Unambiguous OTDR Monitoring in TDM and WDM PON Access Networks". Proc. 11th International Conference on Transparent Optical Networks, ICTON'09, IEEE Xplore, pp. 1-4, June 2009-July 2009.

[6] "ITU-T G.983.1: Broadband optical access systems based on Passive Optical Networks (PON)", January 2005. https://www.itu.int/rec/T-REC-G.983.1-200501-I/en, October 2015.

[7] "ITU-T G.983.2: ONT management and control interface specification for B-PON", July 2005. http://www.itu.int/rec/T-REC-G.983.2-200507-I, October 2015.

[8] "ITU-T G.983.3 A broadband optical access system with increased service capability by wavelength allocation, March 2001. https://www.itu.int/rec/T-REC-G.983.3-200103-I/en, October 2015.

[9] "IEEE P802.3ah Ethernet in the First Mile Task Force, June 2004. http://www.ieee802.org/3/ah/, October 2015.

[10] "ITU-T G.984.1: Gigabit-capable passive optical networks (GPON): General characteristics", March 2003. http://www.itu.int/rec/T-REC-G.984.1-200803-I/en, October 2015.

[11] "ITU-T G.984.2: Gigabit-capable Passive Optical Networks (G-PON): Physical Media Dependent (PMD) layer specification", March 2003. https://www.itu.int/rec/T-REC-G.984.2-200303-I/en, October 2015.

[12] "ITU-T G.984.3:. Gigabit-capable Passive Optical Networks (G-PON): Transmission convergence layer specification", March 2008. http://www.itu.int/rec/T-REC-G.984.3-200803-S, October 2015.

[13] "ITU-T G.984.4: Gigabit-capable passive optical networks (G-PON): ONT management and control interface specification, February 2008. https://www.itu.int/rec/T-REC-G.984.4/en, October 2015

Security Implications of Software Defined Networking in Industrial Control Systems

György Kálmán

mnemonic AS
Oslo, Norway
Email: gyorgy@mnemonic.no

Abstract—**Software-defined Networking (SDN) is appealing not only for carrier applications, but also in industrial control systems. Network engineering with SDN will result in both lower engineering cost, configuration errors and also enhance the manageabiliy of DCS. This paper provides an overview of the applicability of SDN in an control system scenario, with special focus on security and manageability. It also shows the possible enhancements to mitigate the challenges related to network segmentation and shared infrastructure situations.**

***Keywords*–automation, infrastructure, manageability, DCS, SDN**

I. Introduction

Industrial Ethernet is the dominating technology in distributed control systems and is planned to take over the whole communication network from office to the field level, with sensor networks being the only exception at the moment.

Since its introduction in time critical industrial applications, Ethernet's performance has been questioned, mainly because of the old, non-switched networks. Now these problems are solved, automation networks are built with switches, have plenty of bandwidth and the more demanding applications have their specific technologies. These solutions provide intrinsic Quality of Service (QoS), e.g., EtherCAT or try to implement extensions to the Ethernet standards with, e.g., efforts to implement resource reservation like the IEEE 802.1 Time-Sensitive Networking Task Group.

With the industry moving towards Commercial Off The Shelf (COTS) products in the networking solutions (both hardware and software) opened for direct interconnection of other company networks towards the automation systems [4]. This facilitated data exchange in an easier way, but also opened the possibility to attack the previously island-like automation systems from or through the company network [5]. As a result of opening the automation network to be attacked through other systems, a possible categorization of attackers is given by [7]:

- Hobbyists break into systems for fun and glory. Difficult to stop, but consequences are low
- Professional hackers break into systems to steal valuable assets, or on a contract basis. Very difficult to stop, consequences usually financial. May be hired to perform theft, industrial espionage, or sabotage
- Nation-States and Non-Governmental Organizations break into systems to gather intelligence, disable capabilities of opponents, or to cause societal disruption
- Malware automated attack software. Intent ranges from building botnets for further attacks, theft, or general disruption. Ranges from easy to stop to moderately difficult to stop.
- Disgruntled employees, including insider threat and unauthorized access after employment.

Engineering efforts have been made to reduce the risks associated with this interconnection, but it only gained momentum after the more recent incidents of, e.g., stuxnet and repeated cases of Denial of Service (DoS) incidents coming from external networks. The first efforts were focused on including well-known solutions from the IT industry: firewalls, Intrusion Detection Systems (IDS), authentication solutions.

The challenge with these solutions is, that they were designed to operate in a different network environment [6]. Amongst others, the QoS requirements of an automation system tend to be very different than of an office network. The protocol set used is different and the typical protocol inside an automation system runs on Layer 2 networking and not on the IP protocol suite [8].

Beside the efforts on adopting IT security solutions to industrial environments, several working groups are involved in introducing security solutions into automation protocols and protocols used to support an automation system (e.g., IEEE 1588v3 on security functions, IEC 61850 to have integrity protection). The necessity of network management systems are gaining acceptance to support life-cycle management of the communication infrastructure.

In this landscape, Software Defined Networking (SDN) is a promising technology to support automation vendors to deploy their Distributed Control Systems (DCS) more effectively, to allow easier brownfield extensions and to have a detailed overview of the traffic under operation.

The paper is structured as follows: the second section gives a short introduction of Industrial Ethernet and SDN, the third provides an overview of DCS structures, the fourth provides a risk analysis of DCS with SDN, the fifth proposes mitigation solutions for the risks found. The last section draws the conclusion and provides an outlook on future work.

II. Industrial Ethernet and SDN

Industrial Ethernet is built often as a special mixture of a few high-end switches and a large number of small port count discrete or integrated switches composing several network segments defined by both the DCS architecture and location constraints.

Engineering of networks composed from small switches results in typically a magnitude more devices than a comparable office network (e.g., a bigger refinery can have several

ISBN: 978-1-61208-432-9

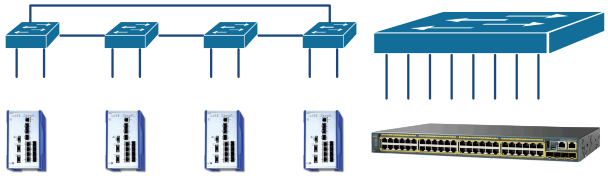


Figure 1. Low port count switches in automation

hundreds of switches with a typical branching factor of 4-7) as shown on Figure 1. The engineering cost and the possibility of configuration-related delays has a big impact on competitiveness.

In the majority of cases, the actual configuration of the devices can be described with setting port-Virtual LAN (VLAN) allocations, Rapid Spanning Tree (RSTP) priorities, Simple Network Management Protocol (SNMP) parameters and performance monitoring [3]. These steps currently require manual work.

SDN is a promising technology in this field, as it has already shown its capabilities for separating traffic and control on carrier networks, the possibility of deploying new services without disturbing the production network and the appealing possibility of having a full overview of network flows from one central controller.

With SDN, a telecom-like network structure is introduced into distributed control systems with splitting the control and the forwarding plane. In such a network, the flows are programmable through a central entity on the control plane. This allows testing and resource reservation for specific flows, not just at commissioning, but also during operation. The ability to isolate new traffic flows can be beneficial from both security and operational viewpoints. These possibilities are appealing for the industrial automation systems, as they are very much in line with the current trends of redundancy, QoS and shared infrastructure.

As defined by the Open Networking Foundation, SDN offers the following features:

- *Directly programmable* Network control is directly programmable because it is decoupled from forwarding functions.
- *Agile* Abstracting control from forwarding lets administrators dynamically adjust network-wide traffic flow to meet changing needs.
- *Centrally managed* Network intelligence is (logically) centralized in software-based SDN controllers that maintain a global view of the network, which appears to applications and policy engines as a single, logical switch.
- *Programmatically configured* SDN lets network managers configure, manage, secure, and optimize network resources very quickly via dynamic, automated SDN programs, which they can write themselves because the programs do not depend on proprietary software.
- *Open standards-based and vendor-neutral* When implemented through open standards, SDN simplifies network design and operation because instructions are provided by SDN controllers instead of multiple, vendor-specific devices and protocols.

SDN architecture is typically represented with three layers (Figure 2). Using several planes in a communication technology is not new, it was present both in ATM, SDH or all the digital cellular networks. What is new, that these management possibilities are now available also in a much smaller scale. It is expected that a network with a centrally managed control plane can better react on changes in traffic patterns and also be more flexible in network resource management. The forwarding performance, however is expected to be very similar or equivalent to the currently used switches, so the industrial applications can run without disturbance in a stable network state.

The normal communication traffic is expected to be significantly larger than the control and signalling traffic generated by SDN and therefore not considered as a performance problem. Typical communication on an industrial network supports the mitigation of this performance threat, as most of the sessions are periodic machine to machine (M2M), which can be scheduled or event driven, with precisely defined transmission deadlines. The gaps between planned periodic traffic are rarely filled with event-driven communication.

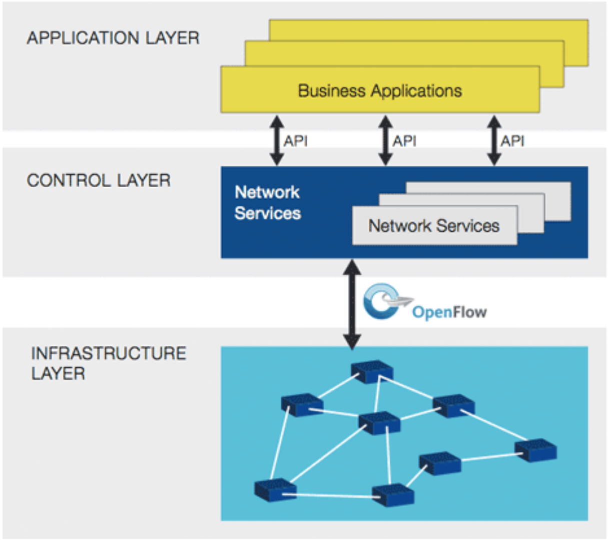


Figure 2. Three layer SDN architecture [12]

III. DCS ARCHITECTURE

Control systems are traditionally built using three network levels. The plant, the client-server and the control network. These levels might have different names, but they share the following characteristics:

- *Plant network* is home of the traditional IT systems, like Enterprise Resource Planning (ERP), office services and other support applications. It is typically under the control of the IT department.
- Client-server network is the non-time critical part of the automation system, where the process-related workplaces, servers and other support entities are located. It is firewalled from the plant network and is under the control of Operations.
- Control network includes everything close to the actual process: controllers, sensors, actuators and other

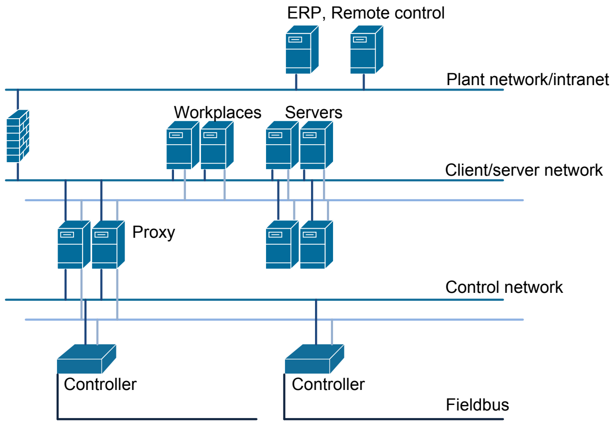


Figure 3. Traditional DCS network architecture

automation components. Typically, it follows a strict time synchronization regime and contains the parts of the network with time-critical components. It is accessible through proxies from the client-server network and under the control of Operations.

IV. Security Landscape

Industrial deployments were built traditionally as isolated islands, thus security was more a question of doors and walls then IT [7]. Employees from the operations department had the responsibility to keep the communication network intact.

Security issues connected to computer networks came with, amongst others, the Supervisory Control and Data Acquisition (SCADA) applications, where remote access to industrial deployments was granted. With the spread of Ethernet and IP-based communication, more and more automation networks could be connected to other networks, to allow easier management and new applications.

Threat analyses showed that industrial systems can be more prone to DoS and related attacks due to the more strict QoS requirements and lack of available processing power in the devices [9]. Typically the deployed network infrastructure can handle a magnitude higher traffic than the end-nodes. This helps in supporting the SDN operation with allowing the traffic, which does not match any of the forwarding rules to be sent to the controller in the normally unused bandwidth. The static traffic picture will also allow the use of sharp heuristics on new traffic, categorizing unknown traffic very early as malicious and drop it early.

DoS attacks require no knowledge of the automation system, only access to the infrastructure, which is a much larger attack surface this case as DCS and especially SCADA systems have a tendency to cover large areas, where enforcing of a security policy (both physical and cyber) is a hard task [10].

This properties have focused the security efforts on protecting the leaves of the network and also on creating policies to ensure the use of hardening practices.

Standard hardening procedures in current industrial deployments include:

- Creation of a *Security Policy* following, e.g., the IEC 62443 standard. This allows to have a structured approach for operating the network.
- A standard way to introduce anti-virus solutions in the automation network using central management.
- Specific focus on the configuration of server and workstation machines with, e.g., policies and additional software components.
- Access and account management: using Role-Based Access Control (RBAC), OS functions like the Group Policy Object (GPO) or tools like a trusted password manager.
- Backup and restoration as a part of disaster recovery.
- Network topology to support security levels in the IEC 62443, with using firewalls as separator.
- Specific remote access solution and whitelisting of both traffic and nodes.

These tasks show that there is an understanding of the importance of security in this field and there are efforts on standardization.

The problematic part of the process is, where these guidelines, policies and physical appliances need to be deployed in a new or an existing installation.

Correctness of the implementation is crucial for future reliability of the system. In a typical current workflow, configuration and deployment of devices is a manual task together with the as-built analysis under or before the Factory Acceptance Test (FAT). At the moment there is no merged workflow and software support for all of the steps mentioned earlier.

SDN can be part of the answer: the communication infrastructure, communication security and monitoring under operation can be implemented using SDN, where the whole or part of the tasks could be automated.

V. SDN-related challenges

SDN changes the security model considerably. To enable automatic features, the operation and the way of controlling a SDN system has to be analysed in the industrial context.

A. The plane structure

After the author's view, the introduction of the separated control and forwarding plane is the biggest enhancement for network security in this relation. In the telecommunication field, separated planes are used since decades to support secure service delivery with minimizing the possibility of a successful attack from the user side towards network management.

In an industrial context, the split planes mean, that the configuration of the devices is not possible from the network areas what clients can see, thus intruders getting access to, e.g., the field network through a sensor, will not be able to communicate with the management interfaces.

Attacks at the data plane could be executed with, e.g., gaining access to the network through a physical or virtual interface and try to execute a DoS attack or a type of fuzzing attack, which might exploit a flaw in the management or automation protocols.

An attacker could also leverage these protocols and attempt to instantiate new flows into the device's forwarding table. The attacker would want to try to spoof new flows to permit specific types of traffic that should be disallowed across the network [18].

B. The SDN controller

The first group of issues are related to the SDN controller. To allow a central entity to control and configure the whole network, it has to gain administrative access over the whole network infrastructure configuration and status. The SDN controller's ability to control an entire network makes it a very high value target.

This can be problematic if the controller has to cross several firewalls to reach all nodes under its control. In the traditional DCS network architecture (Figure 3), in order to gain control of the whole network, the controller has to pass the firewall between the plant and the client-server network, the proxy towards the control network and the controllers towards the field devices.

In a realistic situation, the controller of the DCS will not be allowed to control also the plant network, but is expected to reside inside the DCS, most probably on the client-server network. Inside the automation network, firewalls and the controllers can be configured so, that they pass the SDN signalling.

Network intelligence is being transferred from the network nodes to the central controller entity. This, if being implemented inside a switched network, might only be a semantic difference in network control, as it extends the possibilities of a Network Management System (NMS), but it doesn't need to integrate more sophisticated devices in an industrial situation.

It is expected that a network with a centrally managed control plane can better react on changes in traffic patterns and also be more flexible in network resource management.

In addition to the attack surface of the management plane, the controller has another attack surface: the data plane of the switches. When an SDN switch encounters a packet that does not match any forwarding rules, it passes this packet to the controller for advice. As a result, it is possible for an attacker who is simply able to send data through an SDN switch to exploit a vulnerability on the controller [16].

To mitigate the single-point-of-failure what the SDN controller represents, in most installations, it will be required to deploy two of the controllers in a redundant installation.

Also shared infrastructure between different operators can be a problem in this case. Legal issues might arise if the audit and logging of SDN-induced configuration changes is not detailed enough.

C. Service deployment security

In an SDN case, the controller entity can change the configuration and forwarding behaviour of the underlying devices. This possibility is a valuable addition to the existing set of features, because an SDN system could deploy a new service without disturbing the current operation, which would reduce costs related to scheduled downtimes.

Also, the fine-grained control of network flows and continuous monitoring of the network status offers a good platform for Intrusion Detection Systems (IDS), Managed Security Services (MSS) or a tight integration with the higher operation layers of the DCS.

D. Central resource management

Currently, SNMP-based NMSs are widely used for monitoring the health and status of large network deployments. Using SDN could also here be beneficial, as the monitoring functionality would be extended with the ability of actively changing configurations and resource allocations if needed.

One of the most significant technological and policy challenges in an SDN deployment is the management of devices from different providers. Keeping the necessary complexity and configuration possibilities is hard to synchronize with entities delivered from different providers.

With SDN's abstraction layer one can hide differences in features but also can introduce problems in logging and audit. Network equipment manufacturers are not supporting by default that their devices are managed by a third party.

Although, the rollout of new services would become safer, as the system could check if the required resources are available and the use of SDN is not expected to have a negative impact on the reliability of the network the problems related to shared infrastructure need to be elaborated further.

E. Security implications of shared infrastructure

As part of the universal use of Ethernet communication, it is now common for vendors to share the network infrastructure to operate different parts of an installation. For example, a subsea oil production platform, which is controlled through a hundreds of kilometres long umbilical, can have a different operator for the power subsystem, an other one for the process control and a third one for well control.

In the current operation regimes, the configuration of the networks is rarely changing and all vendors have a stable view of their part of the network shared with the one being the actual operator. With SDN, the network could be controlled in a more dynamic way.

From the technological viewpoint, the biggest challenge is to find a solution, where both the controller and the devices support encrypted control operations. If they support it, than the logging and audit system has to be prepared for a much more dynamic environment.

From a policy management viewpoint, the possibility of fast per-flow configuration opens for new types of problems: the valid network topology and forwarding situation might change fast and frequently, which is not typical in the industry. Logging has to provide the current and all past network configurations with time stamping to allow recreation of transient setups in case of communication errors.

In such a shared case, the use of SDN could reduce risk in topology or traffic changes, as vendors could deploy new services without an impact on other traffic flows in the network. It is possible to create an overlay network, which follows the logical topology of an application or subsystem. This would improve the control possibilities as the staff could follow the communication paths in a more natural way.

F. Wireless integration

Another key field currently is the integration of wireless networks into industrial deployments. SDN could help with integration of wireless technologies by checking if the needs of a new service, e.g., can be satisfied with a path having one or more wireless hops or a new rule has to be deployed into the network to steer the traffic of that service on a different path.

G. Integrating Security in the preliminary design

In the bidding phase, the control engineer could leave the planning of the network on a high level with having an SDN rule set to check if the network can be built. The needed security appliances and other entities would be added to the list of required components following rules developed using the relevant standards.

The control engineer could add the control processes and the SDN software will check if the required resources are available on the communication path. In contrast with current methods, the acceptance of a communication session would also give a proof that the required resources are available and the security requirements are met.

H. Network simulation and capacity estimation

The use of SDN and the central management entities will also lead to more detailed information on network traffic and internal states. The data gathered on operational network not only supports the management of the current network, but also can be used to fine-tune the models used in early steps of bidding and planning and can lead to a more lean approach on network resource allocation. SDN could provide better communication security by helping to avoid overloaded network situations.

I. Firewalls

A current limitation on the coverage of SDN is connected to accountability. While automatic changes in the forwarding table on layer 2 is not expected to cause big problems, automatic rule generation for firewalls and other higher layer devices might cause more problems than it solves.

Granting the control rights of network security devices to the SDN controller is necessary to gain full control over all network nodes. The challenge with this setup is, that L2 forwarding can be described with relative few properties, routing tables with some more, but still within a limited size, firewall rules can contain a lot more properties and values to fill. If automatic generation is disabled, then the SDN network split into several security zones, can only be partially managed by the controller. If automatic generation is enabled, it can cause security breaches (e.g., the early implementations of Universal Plug and Play (UPnP)). This setup also potentially requires cooperation from several companies, e.g., an MSS provider running the security infrastructure and the operations staff at the location focusing on automation.

From the practical viewpoint, there are several issues. The first is that in most cases, management protocols only offer the implementation of security functions, but they are optional, so having a required encryption (one cannot avoid this when managing firewalls) might result in incompatibility already in the communication. The second is, that one needs much more complex support for firewalls in the management software than for switches or routers.

J. Intrusion Detection Systems

Running IDS in an SDN network is promising. It can be the IDS itself, if there is some logic running on the controller.

Current IDS implementations typically use distributed wiretaps or other traffic monitoring sources to watch for malicious traffic and might get aggregated traffic information (e.g., over NetFlow).

SDN can take this functionality into a whole new level. The controller has a complete view of the L2 traffic streams over the whole network, thus not only has a wiretap *everywhere*, but also has the control of the forwarding entities: it can make changes in the forwarding decisions in real time. In extreme cases this can result in, that the malicious packet cannot even travel through the network to its destination, because at the entry the IDS system classifies it as potentially malicious and in transit redirects it into an isolated network.

Industrial deployments are an excellent basis to develop such a fast-reaction IDS: the communication is typically M2M, the network traffic is stationary (whole-new traffic flows are not typical) and the topology is mostly static. The heuristics of the IDS could be as a result, very sensitive on non-planned traffic, thus reacting fast on potential hazards.

If the SDN infrastructure is available because of network management, the extension of providing IDS and firewall management can also lead to cost reduction compared to deploying and operating a separate solution for both.

K. Protecting the SDN controller

As it was mentioned earlier, the SDN controller represents a single-point-of-failure in the network. As most of the industrial deployments are redundant, it is natural to require also a redundant deployment of the SDN controller.

This redundancy is required both from the availability viewpoint (all crucial components have redundant counterparts in most deployments) and also from network security: protection from, e.g., DoS attacks.

Transport security shall be ensured with up to date standard protocols, e.g., TLS for web access or SSH for shell. An effort shall be used to keep the cryptographic suites, which are used by these protocols updated.

VI. Conclusion

SDN is very likely to be the next big step in industrial networks. It offers exactly the functionality automation engineers are looking for: hiding the network and allowing the planning and deployment of network infrastructure without deep technical knowledge, based only on definition of network flows and automatic dimensioning rules.

With a complete view over the current network traffic situation, Quality of Service parameters can be checked in a formal way with the help of the central management entity and as such, provide a proof in all stages of the engineering work, that the infrastructure will be able to support the application.

In brown field extensions, SDN can reduce risks associated with deploying new equipment and extending the current infrastructure because of the isolation of traffic flows and the complete control over the forwarding decisions.

Network security is the other main area, where, if properly planned and implemented, SDN can provide a big step forward in both security and operational excellence. With the real-time overview on the network infrastructure, an SDN-based IDS could react much faster on attacks.

Technological advancements are clearly moving towards a more automated network infrastructure and in the industrial case, SDN is a promising technology, which has to be taken seriously.

REFERENCES

[1] Gy. Kalman, *Applicability of Software Defined Networking in Industrial Ethernet*, in Proceedings of IEEE Telfor 2015, pages 340-343, Belgrade, Serbia

[2] Hirschmann, *Reference Manual, Command Line Interface Industrial Ethernet Gigabit Switch* Release 7.0, Hirschmann, 2011.

[3] A. Gopalakrishnan, *Applications of Software-Defined Networks in Industrial Automation*, https://www.academia.edu/2472112/Application_of_Software_Defined_Networks_in_Industrial_Automation, Accessed 28.05.2015.

[4] N. Barkakati, G. C. Wilshusen, *Deficient ICT Controls Jeopardize Systems Supporting the Electric Grid: A Case Study*, Securing Electricity Supply in the Cyber Age, Springer, pages 129-142, e-ISBN 978-90-481-3594-3

[5] ABB, *Security for Industrial Automation and Control Systems*, White Paper, ABB, Doc. Id. 3BSE032547

[6] Cisco, *Secure Industrial Networks with Cisco*, White Paper, 2015., http://www.cisco.com/c/en/us/products/collateral/se/internet-of-things/white-paper-c11-734453.pdf, Accessed 30.08.2015.

[7] M. McKay, *Best practices in automation security*, White Paper, Siemens, 2012.

[8] C. Alcaraz, G. Fernandez, and F. Carvajal, *Security Aspects of SCADA and DCS Environments*, Critical Infrastructure Protection: Information Infrastructure Models, Analysis, and Defence, LNCS 7130., Springer, September 2012., pp. 120-149

[9] R.C. Parks, E. Rogers, *Best practices in automation security*, Security & Privacy, IEEE (Volume:6 , Issue: 6), 2009., pages 37-43.

[10] I. Fernandez, *Cybersecurity for Industrial Automation & Control Environments*, White Paper, Frost&Sullivan and Schneider Electric, 2013.

[11] D. Cronberger, *The software-defined Industrial Network*, The Industrial Ethernet Book, Issue 84, 2014., Pages 8-13

[12] Open Networking Foundation, *Software-Defined Networking: The New Norm for Networks*, White Paper, https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf, Accessed 28.05.2015.

[13] D. Cronberger, *Software-Defined Networks*, Cisco, 2014, http://www.industrial-ip.org/en/industrial-ip/convergence/software-defined-networks, Accessed 28.05.2015.

[14] R. Millman, *How to secure the SDN infrastructure*, ComputerWeekly, 2015, http://www.computerweekly.com/feature/How-to-secure-the-SDN-infrastructure, Accessed 28.05.2015.

[15] D. Cronberger, *Industrial Grade SDN*, Cisco, 2013, http://blogs.cisco.com/manufacturing/industrial-grade-sdn, Accessed 28.05.2015.

[16] D. Jorm, *SDN and Security*, The ONOS project, 2015, http://onosproject.org/2015/04/03/sdn-and-security-david-jorm/, Accessed 28.05.2015.

[17] G. Ferro, *SDN and Security: Start Slow, But Start*, Dark Reading Tech Digest, 2014, http://www.darkreading.com/operations/sdn-and-security-start-slow-but-start/d/d-id/1318273, Accessed 28.05.2015.

[18] S. Hogg, *SDN Security Attack Vectors and SDN Hardening*, Network World, 2014, http://www.networkworld.com/article/2840273/sdn/sdn-security-attack-vectors-and-sdn-hardening.html, Accessed 28.05.2015.

[19] Open Networking Foundation, *Solution Brief: SDN Security Considerations in the Data Center*, ONF, 2013, https://www.opennetworking.org/images/stories/downloads/sdn-resources/solution-briefs/sb-security-data-center.pdf, Accessed 28.05.2015.

An Authorisation and Access Control Framework for Information Sharing on the Semantic Web

Owen Sacco
DERI, National University of Ireland, Galway
owensacco@deri.org

John G. Breslin
National University of Ireland, Galway
john.breslin@nuigalway.ie

***Abstract*—The Semantic Web brought about open data formats which give rise to an increase in the creation and consumption of structured data. This structured data is easily accessible from SPARQL endpoints, which are considered as the main Web Services in the Semantic Web. Most SPARQL endpoints are publicly available and do not provide fine-grained authorisation and access control enforcement to protect user's personal information. Although a substantial amount of work exists on access control and authorisation on the Web, these cannot be applied directly to structured data due to the different nature of how the data is formatted. In this paper, we present our authorisation and access control framework for Web Services in the Semantic Web. We present several vocabularies that model the different aspects of the authorisation sequence. We also extend our Privacy Preference Manager (PPM) that handles the authorisation sequence for clients accessing the resource owner's personal information in RDF stores.**

Keywords–Access Control, Authorisation, Privacy, Semantic Web.

I. Introduction

The Semantic Web [13] provides formats to enrich information on the Web by annotating Web data with additional meaning that can be processed by machines to offer enhanced services for data sharing and interoperability amongst different data sources. This enables Web agents or Web enabled devices to process this meaning to carry out complex tasks automatically on behalf of users. Moreover, by means of Semantic data, information can be merged easily from heterogeneous sources based on the relationships amongst the data, even if the underlying data schemas differ. The most commonly used meta-formats for the Semantic Web is the Resource Description Framework (RDF) [11]. RDF describes resources on the Web and the relationships between them in the form of *graph models*. RDF uniquely identifies resources on the Web, such as people, events, blog posts, reviews and tags by means of Uniform Resource Identifiers (URIs). Each resource can link to other resources by referring to the URI of the specific resource. The advantage of linking resources is that different datasets can be linked and hence create the *Web of Data*.

RDF data can be queried using SPARQL [9]. SPARQL queries take the form of a set of triple patterns called a basic graph pattern. Most RDF data stores contain a SPARQL Endpoint which accept incoming SPARQL queries over HTTP and return back SPARQL query results. SPARQL Endpoints can be considered as the main Web Services in the Semantic Web. The majority of SPARQL Endpoints follow the RESTful architecture and are publicly accessible.

Web Services allow third-party applications to access and use personal information of end-users. This brought about a need for authorisation mechanisms that allow users to authorise applications to use their information on their behalf. The most common authorisation method is OAuth [3] that provides resource owners to authorise clients to use their protected resources on their behalf without sharing their credentials with the client.

Most SPARQL Endpoints do not use authorisation mechanisms such as OAuth to protect sensitive resources since most SPARQL Endpoints are publicly available. Although the idea behind the Semantic Web is to publish open datasets, this causes a risk to sensitive and personal resources. Several access control models have been proposed, however, most of them do not provide fine-grained authorisation mechanisms for RDF graphs.

In this work, we present our Authorisation and Access Control framework that builds upon our previous work [24]. This framework provides an OAuth architecture for personal information that can be accessed from Web Services in the Semantic Web. Therefore, our framework provides a fine-grained authorisation and access control platform for RDF graphs. Our work is based on the attribute-based access control (ABAC) model since the nature of the Semantic Web is to provide an open environment without knowing *a priori* who will access the data. In this work, we propose several vocabularies to model each authorisation process including credentials and the scope of what a client can access. We also present how this framework is implemented on top of SPARQL Endpoints. Although our main focus for this research paper is on personal data in the Semantic Web, our work can be applied to any use case that requires authorisation for data from Web Services in the Semantic Web.

This paper is structured as follows: Section II provides an overview of some related work. Section III provides an overview of our authorisation and access control framework. In Section IV, we explain how the authentication in this framework functions. Section V presents our vocabularies for describing both the client and the Web Service authorisation component preferences. In Section VI, we explain our vocabularies that model the scope, which provides the necessary permissions for the client (authorised by the user) to access the user's personal information. In Section VII, we explain in detail the authorisation process using the vocabularies described in this paper. Section VIII concludes the paper.

II. Related Work

The authors in [26] propose a method that uses OAuth to authorise clients to access data from triple stores. This work uses usernames and passwords for authentication rather than leveraging the benefits of WebID. Moreover, the OAuth protocol implemented in this work does not use client credentials. Furthermore, the access control ontology presented in this work does not provide fine-grained access control on protected resources but restricts SPARQL clauses. This ontology is also a role-based model that relies on pre-defined access control policies bound to the user's roles.

Similarly, the OpenLink Software Virtuoso [12] RDF store provides OAuth for its SPARQL endpoint. Although this store uses WebID for authentication, the user can authorise or decline the requested SPARQL query rather than fine-grained authorisation for specific resources.

The eXtensible Access Control Markup Language [20] is an XML based language for expressing a large variety of access control policies. Although the XACML is widely used [16], [18], it does not provide the necessary elements to define fine-grained access control statements for structured data. It also does not contain enough semantics to describe what the actual access restriction is about and also does not semantically define which attributes a client or requester must satisfy. Moreover, this method does not provide an authorisation method to authorise third-party applications to use resource owner's data on their behalf.

The Protocol for Web Description Resources (POWDER) [2] is designed to express statements that describe what a collection of RDF statements are about. The descriptions expressed using this protocol are text based and therefore do not contain any semantics that can define what the description states.

The authors in [19] propose a privacy preference formal model consisting of relationships between objects and subjects. The proposed formal model however does not provide any authorisation mechanism for third-party applications to access RDF stores. The authors in [15] propose an access control framework by specifying privacy rules using the Semantic Web Rule Language (SWRL) [1]. This approach also does not provide any authorisation mechanisms for third-party applications to access SPARQL Endpoints. Moreover, this approach relies that the system contains a SWRL reasoner. In [17] the authors propose a relational based access control model called `RelBac` which provides a formal model based on relationships amongst communities and resources. This approach requires to specifically define who can access the resource(s) but does not provide any authorisation mechanism for third-party applications to access SPARQL endpoints.

The authors in [14] propose an expressive logic-based technique for the specification of security properties. However, this approach requires another language and framework in order to process security policies, unlike our work that uses RDF to express the polices that can be processed by the SPARQL engine.

III. Authorisation and Access Control Framework – Overview

The Authorisation and Access Control Framework, illustrated in Figure 1, provides authentication and authorisation mechanisms for RDF data. It is designed to be deployed over SPARQL Endpoints as a Web Service to control and filter data accessed by third-party clients. The authorisation flow in this framework follows the OAuth 2.0 [3] sequence. This Authorisation and Access Control framework is developed within the Privacy Preference Manager (PPM).

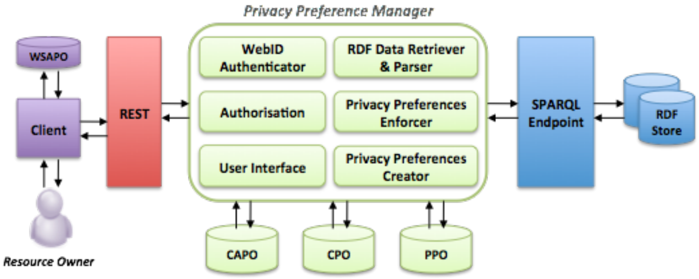


Figure 1. The Authorisation and Access Control Framework

The PPM [22], [24] is an access control manager that provides users to create fine-grained access control policies, known as privacy preferences, for RDF data. The manager also filters requested data by returning back only a subset of the data which is granted access as specified by the privacy preferences. The PPM was developed as a Web application either as a centralised Web application hosted on a centralised server or in a Federated Web environment where users could host their own manager on servers where they desire.

The PPM provides users to login to their manager and create their privacy preferences for their RDF data. Moreover, they could also login to other user's manager and request data. The PPM would return back only that data which the user is granted access – based on the privacy preferences.

The PPM also offers an API which could be integrated within other applications. However, the PPM does not provide any Web Service API whereby third-party applications could call the PPM over HTTPS to benefit from the access control features which it offers. Moreover, it does not provide any authorisation methods to enable users to authorise third-party applications which information they could access and consume. Therefore, we have extended the PPM to provide RESTful methods where third-party applications could send their SPARQL query to the manager and receive back the filtered RDF data. Furthermore, we have extended the PPM with an authorisation component to handle the authorisation process of third-party applications.

The PPM is designed to handle the requests sent to SPARQL Endpoints from client applications using the RESTful architecture. The PPM handles the requests by (1) requesting the resource owner (i.e. user) to authenticate with the PPM; (2) requesting the resource owner to authorise the client which data it can consume; and (3) sends back a filtered subset of the data which the client is authorised to access. The SPARQL Endpoints should be configured to accept requests sent only from the PPM.

The authentication and authorisation sequence in our framework, as illustrated in Figure 2, consists of: (1) the resource owner (i.e. user) requests a service from the client; (2) the client sends a request for temporary credentials to the PPM – the request includes the client credentials that identify

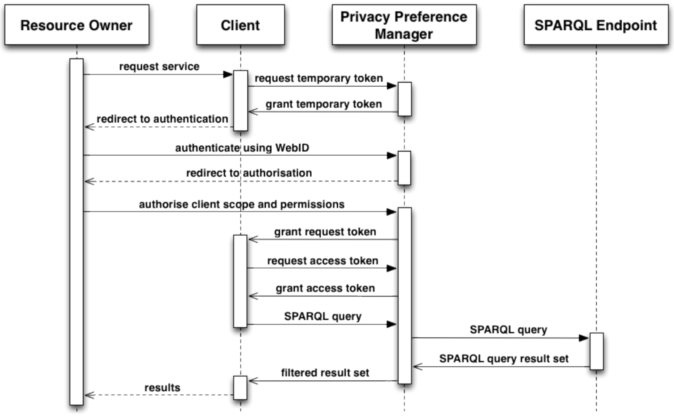


Figure 2. The Authorisation and Access Control Sequence

the client, and the temporary credentials identify the authorisation sequence; (3) the temporary credentials are granted to the client; (4) the client redirects the resource owner to the PPM – the redirect request includes the client's callback URI and the temporary credentials; (5) the resource owner authenticates with the PPM; (6) the resource owner authorises the client by selecting which scope and permissions that will be granted to the client; (7) the PPM sends back the temporary token including a verifier to the client; (8) the client then exchanges the verified temporary token to the access token credentials by sending a request to the PPM – the request includes the temporary token and the verifier; (9) the PPM sends back the access token credentials to the client; (10) the client then sends the SPARQL query (together with the access token credentials) to the PPM; (11) the PPM sends the SPARQL query to the SPARQL Endpoint and the SPARQL Endpoint sends back the query result; (12) the PPM sends back the client only a filtered result set based on what the resource owner had authorised the client which data it can access; and (13) the client renders the service and displays the results to the resource owner.

IV. AUTHENTICATION

Most Web applications request users to provide a username and password in order to authenticate themselves into the system. In Semantic Web applications, the WebID protocol [25] is used as an authentication method. It provides a mechanism whereby users can authenticate using FOAF and X.509 certificates over SSL. The digital certificates contain the public key and a URI that points to the location where the FOAF profile is stored. The WebID authentication mechanism parses the WebID URI from the certificate and retrieves the FOAF profile from its location. The public key in the certificate and the public key in the FOAF profile are checked to grant the user access if the public keys match. The WebID certificates can be self-signed certificates.

Once the resource owner is redirected to authenticate with the PPM, the resource owner is requested to provide a WebID certificate. The PPM's authentication module handles the WebID authentication process by using a WebID verifier that checks that the keys in both the certificate and the FOAF profile match. The advantage of using URIs to identify resource owners is that it eliminates the users to register or create multiple accounts on various servers. If the keys match, then the resource owner is authenticated with the PPM.

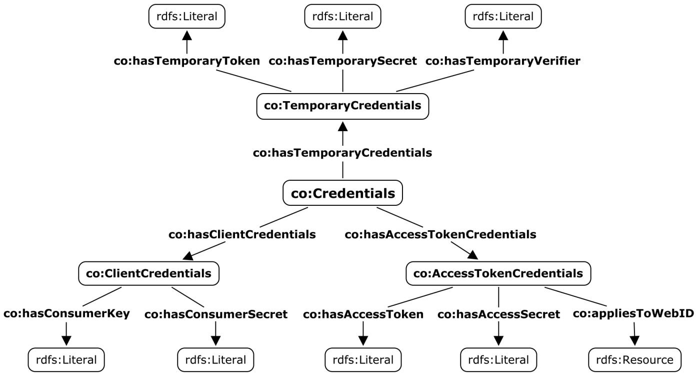


Figure 3. Credentials Ontology (CO)

Definition 1: Authentication. Let PPM be a PPM instance, $Cert$ an SSL digital signed certificate, O a resource owner identified by a URI and P a resource owner's FOAF profile. Let $Certificate(Cert, O)$ mean that $Cert$ is the SSL certificate of O, $Profile(P, O)$ mean that P is the profile of O, $Verify(Cert, P)$ mean that the public key in $Cert$ is verified with the public key in P and $Authenticate(PPM, O)$ mean that O is authenticated with PPM. Thus, Authentication is defined:

$$\begin{aligned} &\textit{Certificate(Cert,O)} \wedge \textit{Profile(P,O)} \wedge \textit{Verify(Cert,P)} \\ &\qquad \Rightarrow \textit{Authenticate(PPM,O)} \end{aligned} \tag{1}$$

V. MODELLING AUTHORISATION PREFERENCES

In this section we present: (1) the *Credentials Ontology (CO)* which is a light weight vocabulary to describe both the client and the Web Service (i.e. the PPM) credentials; (2) the *Client Authorisation Preferences Ontology (CAPO)* which is a light weight vocabulary to describe the client details when a client registers with the PPM; and (3) the *Web Service Authorisation Preference Ontology (WSAPO)* which is a light weight vocabulary to describe the details of the Web Service authorisation component (i.e. the PPM) to be used by the client during the authorisation process.

A. Credentials Ontology (CO)

The *Credentials Ontology (CO)* [7], illustrated in Figure 3, is a light weight vocabulary to describe three types of credentials: (1) *temporary or request token credentials*; (2) *client credentials*; and (3) *access token credentials*.

The *temporary or request token credentials* identify an authorisation sequence. These tokens are randomly generated for each authorisation request. The *client credentials* identify a particular client. These credentials are created when a client registers with a PPM in order to be able to access the data stored within the SPARQL Endpoint. The *access token credentials* are generated by the PPM each time after the resource owner authorises the client to use his/her personal data on his/her behalf. The access token credentials identify the scope and permissions which the resource owner granted the client at a particular instance.

The *Credentials Ontology (CO)* provides the following classes and properties to describe the three types of credentials:

- `co:Credentials` is the main class of CO and the credentials described using this vocabulary will be instances of this class.
- `co:TemporaryCredentials` is a class that describes the temporary or request token credentials. This class provides the `co:hasTemporaryToken` property that defines an identifier for an authorisation request. This identifier is generated whenever the client requests an authorisation sequence. The `co:hasTemporarySecret` property defines the shared secret generated by the PPM for the authorisation request. This shared secret is used for signing the authorisation requests. This class also provides a `co:hasTemporaryVerifier` property that defines a verification identifier generated by the PPM once the resource owner authenticates and completes the authorisation process.
- `co:ClientCredentials` is a class that describes the client credentials. This class provides the `co:hasConsumerKey` property that defines an identifier for a client sending requests to the SPARQL Endpoint through the PPM. This identifier is generated when the client registers with the PPM to consume the data from the SPARQL Endpoint. Therefore, the client must store this identifier and use it whilst sending requests to the PPM. This class also provides the `co:hasConsumerSecret` property that defines the shared secret generated by the PPM. This shared secret is also generated when the client registers with the PPM and it is used for signing the requests. Similar to the consumer key, the client must store this shared secret.
- `co:AccessTokenCredentials` is a class that describes the access token credentials. This class provides the `co:hasAccessToken` property which describes the identifier to the client's authorised scope and permissions authorised by the resource owner. This class also provides `co:hasAccessSecret` property which describes the shared secret for signing the requests after the authorisation process is complete. Both the `access token` and the `access secret` are generated by the PPM after the resource owner completes the authorisation sequence. This class also provides `co:appliesToWebID` property which links the access token credentials to the resource owner's WebID URI who authorised the client.

B. Client Authorisation Preferences Ontology (CAPO)

The *Client Authorisation Preferences Ontology (CAPO)* [5], illustrated in Figure 4, is a light weight vocabulary that describes client details which are stored in the CAPO repository – as illustrated in Figure 1. These details are created once the client is registered with the Web Service (i.e. the PPM). The client details are used by the PPM to verify clients during the authorisation process.

The CAPO vocabulary provides the following classes and properties:

- `capo:Client` is the main class of CAPO and instances of this class define clients that can make use of the authorisation sequence of the Web Service (i.e. the PPM).

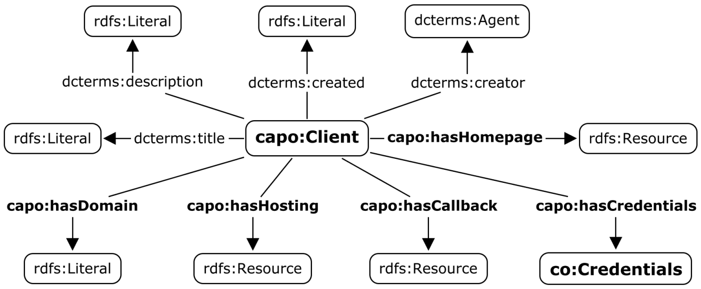


Figure 4. Client Authorisation Preferences Ontology (CAPO)

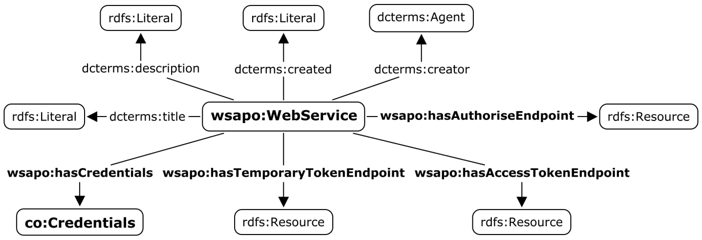


Figure 5. Web Service Authorisation Preferences Ontology (WSAPO)

- `capo:hasDomain` is a property that defines the client's domain. This is used as additional security to allow requests received only from this domain.
- `capo:hasHosting` is a property that defines the URI where the client is hosted on.
- `capo:hasCallback` this property defines the client's callback URI. Although the callback URI is passed within the requests, this is also used for additional security since the callback URI in the request must match the callback URI defined using this vocabulary.
- `capo:hasCredentials` this property defines the client's credentials defined using the Credentials Ontology (CO) which are generated on registration with the Web Service (i.e. the PPM).
- `capo:hasHomepage` this property defines the client's homepage.

Other terms could be used from other vocabularies to define other details such as `dcterms:title` defines the title given to a client; `dcterms:description` defines the client's description; `dcterms:created` defines the date when the client's details were registered; and `dcterms:creator` defines the creator of the client's details.

C. Web Service Authorisation Preferences Ontology (WSAPO)

The *Web Service Authorisation Preferences Ontology (WSAPO)* [10], illustrated in Figure 5, is a light weight vocabulary that describes the details of the Web Service authorisation component which are stored in the WSAPO repository – as illustrated in Figure 1. These details are used by the client during the authorisation process.

The WSAPO vocabulary provides the following classes and properties:

- `wsapo:WebService` is the main class of WSAPO and instances of this class define Web Services that provide the authorisation architecture such as the PPM.

- `wsapo:hasCredentials` this property defines the client's credentials defined using the Credentials Ontology (CO) which are generated on registration with the Web Service (i.e. the PPM). These are used to identify the client during the authorisation sequence.
- `wsapo:hasTemporaryTokenEndpoint` is a property that defines the Web Service's temporary token credentials endpoint. This allows a client to request for temporary token credentials.
- `wsapo:hasAccessTokenEndpoint` is a property that defines the Web Service's access token credentials endpoint. This allows a client to exchange verified temporary token credentials to access token credentials.
- `wsapo:hasAuthoriseEndpoint` is a property that defines the Web Service's authorisation endpoint. This allows a client to use the authorisation architecture by sending the temporary credential tokens to this endpoint. Once the authorisation is complete, the Web Service will return verified temporary token credentials (i.e. the temporary token credentials together with the verifier).

Other terms could be used from other vocabularies to define other details such as `dcterms:title` defines the title given to a Web Service; `dcterms:description` defines the Web Service's description; `dcterms:created` defines the date when the Web Service authorisation component details were created; and `dcterms:creator` defines the creator of the Web Service authorisation component details.

VI. Modelling Permissions

Apart from modelling the details of both the client and the Web Service (i.e. the PPM), the authorisation scope and permissions which the resource owner grants the client in order to access the protected resources should be modelled as well. The scope and permissions are modelled using the Client Permissions Ontology (CPO) – explained in this section. This light weight vocabulary uses the Privacy Preference Ontology (PPO) to model the permissions.

A. Privacy Preference Ontology (PPO) – Overview

The Privacy Preference Ontology (PPO) [8], [21], [23] - is a light-weight Attribute-based Access Control (ABAC) vocabulary that allows users to describe fine-grained privacy preferences for restricting or granting access to non-domain specific Linked Data elements, such as Social Semantic Data. Among other use-cases, PPO can be used to restrict part of FOAF profiles records only to clients or users that have specific attributes. It provides a machine-readable way to define settings such as "Provide my personal phone number only to my family" or "Grant write access to my technical blog only to my co-workers".

As PPO deals with RDF(S)/OWL data, a privacy preference, defines: (1) the resource, statement, named graph, dataset or context it must grant or restrict access to; (2) the conditions refining what to grant or restrict (for example defining which resource as subject or object to grant or restrict); (3) the access control privileges; and (4) a SPARQL query, (`AccessSpace`) *i.e.* a graph pattern representing what must be satisfied by the client or user requesting information. The access control type includes the `Create`, `Read` and `Write` (which also includes `Update`, `Delete` and `Append`) access control privileges.

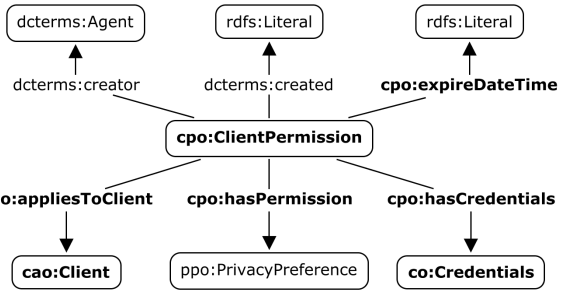


Figure 6. Client Permissions Ontology (CPO)

B. Client Permissions Ontology (CPO)

The *Client Permissions Ontology (CPO)* [6], illustrated in Figure 6, is a light weight vocabulary that describes the scope and permissions which the resource owner grants to the client. The scope and permissions are used by the PPM to grant (or deny) the client access to the resource owner's protected resources.

The CPO vocabulary provides the following classes and properties:

- `cpo:ClientPermission` is the main class of CPO and instances of this class define the scope and permissions the resource owner has granted a particular client.
- `cpo:appliesToClient` this property defines which client (as described using the CAPO vocabulary) the scope and permissions apply to.
- `cpo:hasPermission` is a property that defines the scope and permissions defined using the PPO vocabulary. For example, if the client wants to have access to a particular resource, for instance an email address, the `cpo:hasPermission` would define a `ppo:PrivacyPreference` that would `ppo:appliesToResource` the email address with an `acl:Read` access control privilege.
- `cpo:hasCredentials` this property defines the temporary token credentials and the access token credentials defined using the Credentials Ontology (CO) once these are generated by the PPM and granted to the client.
- `cpo:expireDateTime` this property defines when the scope and permissions expire.

Other terms could be used from other vocabularies to define other details such as `dcterms:created` defines the date when the scope and permissions were created and `dcterms:creator` defines the creator.

VII. Authorisation

Whenever the resource owner requests a service from the client, the client reads the temporary token endpoint URI from the WSAPO datastore for that particular Web Service (i.e. PPM). The client sends a request for the temporary token credentials from this endpoint URI and once retrieved, the client redirects the resource owner to authenticate with the PPM.

Once the resource owner is authenticated using WebID as explained in section IV, the PPM first checks within the CPO datastore whether there are any valid access token credentials already granted to that client by that resource owner for the same request. If valid access token credentials exist, the temporary token credentials are verified and sent to the client. Moreover, the client's permissions defined using CPO are created containing the verified temporary token credentials, the access token credentials that already exist and the permissions which were already granted. Otherwise, the PPM checks if there are any privacy preferences in the PPO store created by the resource owner that authorise the client access to the protected resources. If privacy preferences exist, then the client's permissions defined using CPO are created that link to these privacy preferences. The temporary token credentials are also verified and sent to the client.

When neither any valid access token credentials or privacy preferences exist, then the resource owner is presented with an authorisation page whereby the PPM requests the user to authorise the client's request. The requested SPARQL query is first parsed using the ARC2 [4] SPARQL query parser and presented to the resource owner. The resource owner either authorises the client the whole request; or selects which protected resources the client can access; or denies the whole request. Moreover, the resource owner selects the temporality of the permissions by specifying the expiry date and time. However, any authorised credentials can be revoked any time. Depending on the resource owner's decision, the client's permissions are defined using CPO and the temporary token credentials are verified. The client then exchanges the verified temporary token credentials to access token credentials by requesting the access token endpoint URI.

Whenever the client sends the SPARQL query together with the access token credentials to the PPM, the PPM will send back only what the client is granted to access; based on the client's permissions defined using CPO.

Definition: Authorisation. Let C be a client, O a resource owner identified by a URI, R a resource and A an access control privilege. Let $Request(C, R)$ mean that C requested R, $Resource(R, O)$ mean that R is the resource of O, $Assign(A, O)$ mean that A is assigned by O, $AssignAccess(R, A)$ mean that R is assigned access A and $Authorise(R, C)$ mean that C is authorised R. Thus, Authorisation is defined:

$$\begin{aligned}&\textit{Request(C,R)} \wedge \textit{Resource(R,O)} \wedge \textit{Assign(A,O)}\\ &\quad \wedge \textit{AssignAccess(R,A)} \Rightarrow \textit{Authorise(R,C)}\end{aligned} \tag{2}$$

VIII. Conclusion and Future Work

SPARQL endpoints, which are the most commonly used Web Services in the Semantic Web, are publicly accessible and do not provide any authentication, authorisation and access control functionality. Therefore, in this paper we have presented our authorisation and access control framework that provides resource owners to authorise third-party applications to consume their resources within RDF stores on their behalf. We have presented several vocabularies, namely: (1) the *Credentials Ontology (CO)*; (2) the *Client Authorisation Preferences Ontology (CAPO)*; (3) the *Web Service Authorisation Preferences Ontology (WSAPO)*; and (4) the *Client Permissions Ontology (CPO)* that model several aspects of the authorisation sequence. We have also extended the *Privacy Preference Manager (PPM)* to handle the authorisation sequence for SPARQL endpoints.

As future work, we will enhance the PPM to assert the trustworthiness of third-party applications. The authorisation sequence will become more autonomous since clients will have to satisfy a trust value threshold in order to be authorised to consume the resource owner's protected resources.

Acknowledgment

This work is funded by the Science Foundation Ireland under grant number SFI/08/CE/I1380 (Líon 2) and by an IRCSET scholarship co-funded by Cisco systems.

References

[1] SWRL: A Semantic Rule Language Combining OWL and RuleML. http://www.w3.org/Submission/SWRL, 2004. [Online; accessed 31-July-2015].

[2] Protocol for Web Description Resources (POWDER). http://www.w3.org/TR/powder-dr, 2009. [Online; accessed 31-July-2015].

[3] OAuth 2.0 Authorization Framework. http://tools.ietf.org/html/rfc6749, 2012. [Online; accessed 31-July-2015].

[4] ARC 2 RDF Store. https://github.com/semsol/arc2, 2013. [Online; accessed 31-July-2015].

[5] Client Authorisation Preferences Ontology (CAPO). http://vocab.deri.ie/capo#, 2013. [Online; accessed 31-July-2015].

[6] Client Permissions Ontology (CPO). http://vocab.deri.ie/cpo#, 2013. [Online; accessed 31-July-2015].

[7] Credentials Ontology (CO). http://vocab.deri.ie/co#, 2013. [Online; accessed 31-July-2015].

[8] Privacy Preference Ontology (PPO). http://vocab.deri.ie/ppo#, 2013. [Online; accessed 31-July-2015].

[9] SPARQL Query Language for RDF. http://www.w3.org/TR/sparql11-overview, 2013. [Online; accessed 31-July-2015].

[10] Web Service Authorisation Preferences Ontology (WSAPO). http://vocab.deri.ie/wsapo#, 2013. [Online; accessed 31-July-2015].

[11] Resource Description Framework (RDF). https://www.w3.org/RDF, 2014. [Online; accessed 31-July-2015].

[12] OpenLink Software Virtuoso Universal Server. http://virtuoso.openlinksw.com, 2015. [Online; accessed 31-July-2015].

[13] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284:34–43, 2001.

[14] E. Bertino, F. Buccafurri, E. Ferrari, and P. Rullo. A logic-based approach for enforcing access control. *J. Comput. Secur.*, 8(2,3):109–139, Aug. 2000.

[15] B. Carminati, E. Ferrari, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. A Semantic Web Based Framework for Social Network Access Control. In *Proceedings of the 14th ACM Symposium on Access Control Models and Technologies, SACMAT '09*, 2009.

[16] S. Franzoni, P. Mazzoleni, S. Valtolina, and E. Bertino. Towards a fine-grained access control model and mechanisms for semantic databases. In *ICWS 2007*, 2007.

[17] F. Giunchiglia, R. Zhang, and B. Crispo. Ontology Driven Community Access Control. *Trust and Privacy on the Social and Semantic Web, SPOT'09*, 2009.

[18] R. Hebig, C. Meinel, M. Menzel, I. Thomas, and R. Warschofsky. A web service architecture for decentralised identity- and attribute-based access control. In *ICWS 2009.*, 2009.

[19] P. Kärger and W. Siberski. Guarding a Walled Garden Semantic Privacy Preferences for the Social Web. *The Semantic Web: Research and Applications*, 2010.

[20] Oasis. eXtensible Access Control Markup Language (XACML) Version 3.0. 2009.

[21] O. Sacco and J. G. Breslin. PPO & PPM 2.0: Extending the privacy preference framework to provide finer-grained access control for the web of data. In *I-SEMANTICS '12*, 2012.

[22] O. Sacco and A. Passant. A Privacy Preference Manager for the Social Semantic Web. In *Semantic Personalized Information Management Workshop*, SPIM'11, 2011.

[23] O. Sacco and A. Passant. A Privacy Preference Ontology (PPO) for Linked Data. In *Linked Data on the Web Workshop*, LDOW'11, 2011.

[24] O. Sacco, A. Passant, and S. Decker. An Access Control Framework for the Web of Data. In *IEEE TrustCom-11*, 2011.

[25] H. Story, B. Harbulot, I. Jacobi, and M. Jones. FOAF + SSL : RESTful Authentication for the Social Web. *Semantic Web Conference*, 2009.

[26] D. Tomaszuk and H. Rybiński. OAuth+UAO: A Distributed Identification Mechanism for Triplestores. In *ICCCI*, 2011.

Facial Part Effects Analysis using Emotion-evoking Videos: Smile Expression

Kazuhito Sato
Department of Machine Intelligence and Systems Engineering,
Faculty of Systems Science and Technology, Akita Prefectural University
Yurihonjo, Japan
e-mail: ksato@akita-pu.ac.jp

Momoyo Ito
Institute of Technology and Science,
Tokushima University
Tokushima, Japan
e-mail: momoito@is.tokushima-u.ac.jp

Hirokazu Madokoro
Department of Machine Intelligence and Systems Engineering,
Faculty of Systems Science and Technology, Akita Prefectural University
Yurihonjo, Japan
e-mail: madokoro@ akita-pu.ac.jp

Sakura Kadowaki
Smart Design Corp.

Akita, Japan
e-mail: sakura@smart-d.jp

***Abstract*-This study specifically examines the expressive process of "happiness" related facial expressions after giving a stress stimulus. In addition, it presents a quantitative analysis of expressive tempos and rhythms using mutual information. By acquiring image datasets of facial expressions under states of pleasant–unpleasant stimulus for 20 participants, we calculated the information in three region of interests (ROIs): ROI 1, the whole face and the upper face; ROI 2, the whole face and the lower face; and ROI 3 between the upper face and the lower face. Additionally, we tried to express complexity and ambiguity objectively during facial expressions because of human psychological states. Results clarified the possibility of estimating the impression of facial expressions from the magnitude relation and order relation of mutual information of each ROI. More than male participants, female participants were able to create facial expressions of "happiness" easily and intentionally, and were less susceptible to discrepancy expressions.**

Keywords–Psychological measures, stress; Intentional facial expression; Machine learning approaches; Behavior modeling

I. INTRODUCTION

Attractive smiles attract people and represent a symbol of happiness, soothing another person’s mind. Smiles are therefore effective as a lubricant of human communication. According to a study [1] that analyzed geometric features with respect to charming smiles, the most attractive part of smiles in both men and women is perceived as the eye, followed by the mouth. In addition, facial parts associated with the eyes and mouth, such as the corners of the eyes and mouth, are reportedly more important as attractive factors of smiles. In attractive smiles, the existence of a golden ratio was observed in the aspect ratio of the expression rectangle. Furthermore, Yamada et al. [2] investigated the relevance between the whole and partial impression formed from facial parts and pointed out the following points. Eyes play an extremely important role in forming impressions of others. It is possible to some degree to illustrate the overall impression by adding and coupling the partial impression formed from each part. However, they suggest that individual differences exist in the information of the parts which are expected to be related to emphasis. Assessing male and female viewpoints of smile expressions specifically, women are said to tend to expose smiles more than men [3]. Moreover, smiles are natural for women: women are better at making smiles than men. Particularly, women have excellent skills to adjust positive emotional expressions. Such natural expressions can elicit positive effects on a person viewing the smile (recipient). Nevertheless, for the creation of intentional facial expressions, different facial muscles are said to move in conjunction with natural facial expressions [4]. Particularly examining the expressive process, the deformation degree, and operation timing of facial parts creating smiles are expected to vary slightly.

To clarify the relevance between facial expressions and psychological states to date, as a result of verifying the relevance between psychological stress and facial expressions using the framework of Facial Expression Spatial Charts (FESCs), we demonstrated that the degree of stress accumulation can be easily ascertained from facial expression types and expressive processes [5] [6]. Additionally, we proposed a framework of rhythms and tempos that specifically examines actions to repeat intentional facial expressions after giving a stress stimulus [7]. We define one rhythm as one tempo repeated several times. In addition, regarding one tempo as the period during which facial expressions transform from a neutral face (i.e., expressionless) to the next neutral face, we found that the variation in unpleasant stimulus became greater than that in pleasant stimulus, addressing the variation of the number of frames constituting one tempo. Furthermore, using Bayesian networks, we constructed a graphical model of the relation between these three facial expressions and psychological stress factors. Results show that facial expressions displaying the effects of psychological stress easily were "happiness" and "sadness." Additionally, we showed the possibilities that facial parts (such as the eyes and mouth) easily differed by facial expression type [8] [9] [10].

In this study, particularly addressing the expressive process of "happiness" facial expression after giving a pleasant–unpleasant stress stimulus by emotion-evoking videos, we strove objectively to express complexity and ambiguity through facial expression because of human psychological states, by quantitative analysis of expressive rhythms from the viewpoint of mutual information.

This paper is presented as follows. We review related work to clarify the position of this study in Section II. Section III presents a definition of a new framework of exposed rhythms and tempos for analyzing relations of psychological stress and facial expressions. Section IV describes a method to capture facial expression images, in addition to preprocessing, classification of facial expression patterns with self-organizing maps, integration of facial expression categories with fuzzy adaptive theory, and quantification of expressive rhythms using mutual information. We explain our originally developed facial expression datasets including stress measurements in Section V. In Section VI, based on the calculation results of mutual information in a time-series change of ELs for each facial region, we analyze the respective trends exhibited by men and women. Additionally, we discuss the effects of a pleasant–unpleasant stimulus which would give the expressive rhythm of facial expressions from the perspective of mutual information. Finally, we present conclusions and intentions for future work in Section VII.

II. Related Works

In spite of increasing or decreasing attractiveness of a "smile" with changes in expressive process, many conventional studies have examined the shape of a post-expression face. Case studies examining the expressive process are few [11]–[14]. Regarding impression formation of friendly and thoughtful smile expressions, Ishi et al. [11] described the following. A continuous video presentation, such as expression levels from a neutral face become the maximum, is the most effective. Hanibuchi et al. [12] proposed a smile training method that specifically examines facial expressions process. Through impression evaluation experiments, they demonstrated the validity of goal setting with the actor's perspective. In addition, particularly addressing a natural smiling face, Fujishiro et al. [13] [14] investigated how eye, cheek, and mouth movements contribute to the impression formation of natural smiles in the expressive process. Results revealed moderate correlation between the behavioral termination of the eyes and cheek and the impression formation of natural smiles. Nevertheless, the authors did not report the psychological state of the actor when viewing a "natural smile" and "forced smile," such as a disagreement expression or expression suppression. Particularly, they were unable to come up to address impression formation based on the timing structure of facial parts.

For a good impression on the face of a conversation partner, Kampe et al. [15] revealed that the good impression was more emphasized with matching of each other's eye-gaze. Using anthropomorphic agents, Kuroki et al. [16] indicated the following. The combination of eye-gaze and facial expressions affects emphases of impressions. The impressive transmission of friendship properties can be emphasized particularly. Moreover, by analyzing brain activities using functional magnetic resonance imaging (fMRI) as physiological indices, an activation is observed in the prefrontal cortex responsible for higher cognitive functions such as emotional processing, motivation, and reasoning. Furthermore, the same activation is observed in the amygdala associated with emotions and rewards. Therefore, the formation of a good impression shows that the prefrontal cortex and the amygdala play mutually important roles [17]. However, impression evaluation has not been done subjectively for overall impressions of the face. Moreover, dealing with impression formation based on the timing structure of facial parts has not been achieved.

III. Framework of Exposed Rhythms and Tempos

As an index for quantifying individual facial expression spaces, we proposed a framework of expression levels (ELs) [5]. The ELs include both features of the pleasure and arousal dimensions based on the arrangement of facial expressions on Russell’s circumplex model [18]. Specifically, we extract the dynamics of topological changes of facial expressions of facial components such as the eyes, eyebrows, and mouth. Topological changes show the structure defining the connection form of the elements in the set. The ELs obtained in this study are sorted to categories according to their topological changes in intensity from expressions that are regarded as neutral facial expressions. As discussed above, the ELs in this study include features of both pleasure and arousal dimensions. In Russell’s circumplex model, all emotions are constellated on a two-dimensional space: the pleasure dimension of pleasure–displeasure and arousal dimension of arousal–sleepiness. In

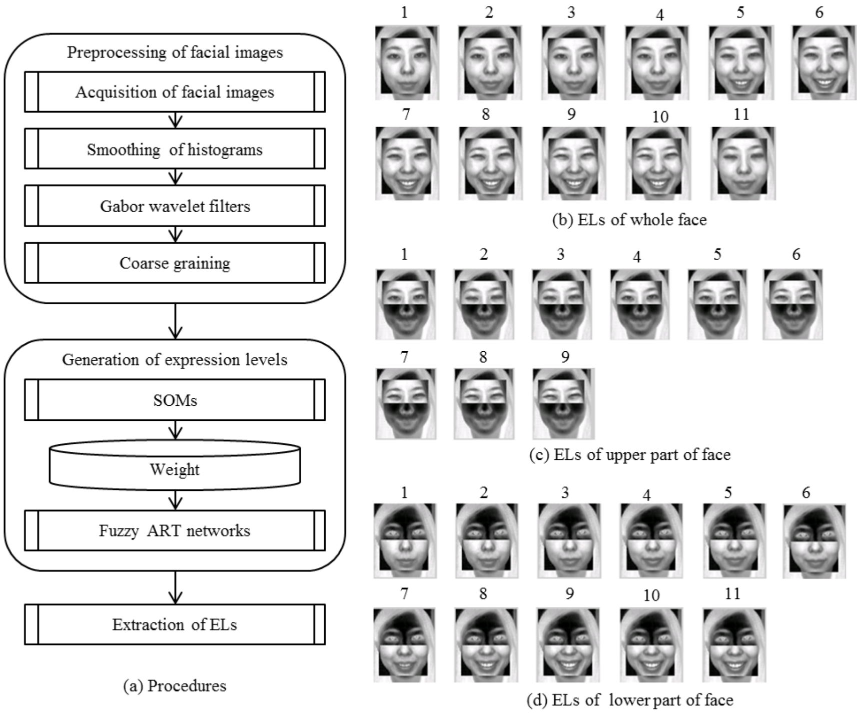


Figure 1. Overview of the procedures used for our proposed method.

the intentional facial expressions covered in this study, direct handling of the facial expressions for the influence of pleasure dimension is difficult. Therefore, as a method of measuring transitory stress response, we conduct an evaluation using the salivary amylase test. As a method of measuring the transitory stress response, we conduct an evaluation using the salivary amylase test during the task of watching emotion-evoking videos causing a pleasant–unpleasant state. Specifically examining the values of salivary amylase activity before and after watching videos, we can effectively perform stress measurements using salivary amylase tests to assess the stress state transiently. Consequently, we target the intentional facial expressions under pleasant and unpleasant stimulation states.

In this study, using temporal variation of ELs, we intend to visualize rhythms and tempos of facial expressions that humans create. We defined one rhythm as a tempo that is repeated several times. One tempo is the period during which facial expressions are transformed from a neutral state to the next neutral state. Facial expressions exhibited intentionally by humans form an individual space based on the dynamic diversity and static diversity of the human face. Facial expression dynamics can be regarded as "topological changes in time-sequential facial expression patterns that facial muscles create." Static diversity is individual diversity that is configured by the facial component position, size, and location, consisting of the eyes, nose, mouth, and ears. In contrast, dynamic diversity denotes that a human can move facial muscles to express internal emotions unconsciously and sequentially or to express emotions as a message. After organizing and visualizing topological changes of face patterns by ELs, we attempt to use the framework of rhythms and tempos with expressions to examine ambiguities and complexities of facial expressions attributable to a psychological state.

IV. Proposed Method

Facial expression processes differ among individuals. Therefore, adaptive learning mechanisms are necessary for modification according to individual characteristic features of facial expressions. In this study, our target is intentional facial expressions. We use self-organizing maps (SOMs) [19] to extract topological changes of facial expressions and for normalization with compression in the direction of the temporal axis. After classification by SOMs, facial images are integrated using Fuzzy ART [20], which is an adaptive learning algorithm with stability and plasticity. In fact, SOMs perform unsupervised classification input data into a mapping space that is defined preliminarily. In contrast, Fuzzy ART performs unsupervised classification at a constant granularity that is controlled by the vigilance parameter. Therefore, using SOMs and Fuzzy ART, time-series datasets showing changes over a long term are classified using a certain standard. Figure 1 presents an overview of the procedures used for our proposed method. In the following, we describe extraction of time-sequential

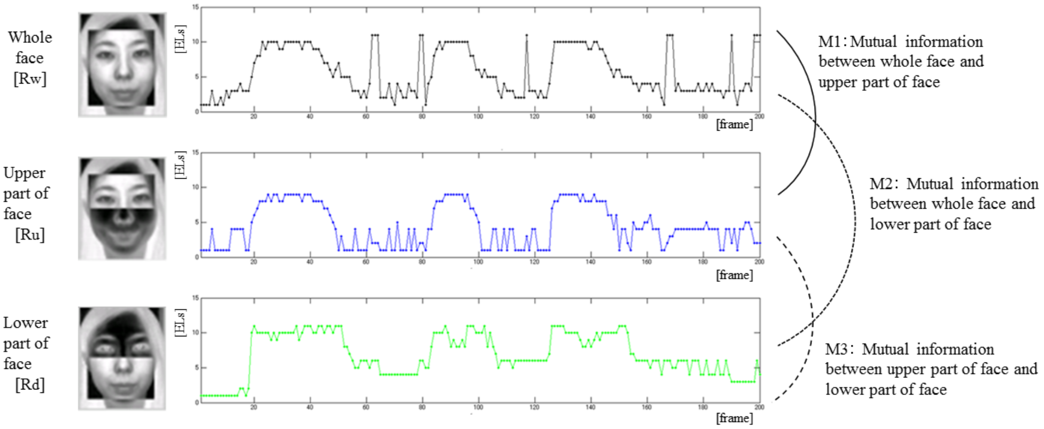


Figure 2. Each mutual information among time-series changes of facial parts.

changes of ELs, and also explain quantification of expressive tempos and rhythms by mutual information.

A. Acquisition of Time-series Variation of ELs

We set the region of interest (ROI) to 90 × 80 pixels, including the eyebrows, which all contribute to the impression of a whole face as facial feature components. With preprocessing, brightness values are normalized for time-series images of facial expressions. The influence of brightness values attributable to illumination conditions is thereby reduced. Moreover, smoothing the histogram is useful to adjust contrast and clarify the images. In addition, using the orientation selectivity of Gabor Wavelets filtering as a feature representation method, the facial parts characterizing the dynamics of facial expressions are emphasized, such as the eyes, eyebrows, mouth, and nose. By down-sampling (i.e., 10 × 10 pixels) time-series facial expressions converted with Gabor Wavelets filtering [21], the effects of a slight positional deviation when taking facial images were minimized. Then data size compression was conducted.

First, SOMs are used to learn the time-series images of facial expressions with down-sampling. The face images showing topological changes of facial expressions that are similar are classified into 15 mapping units of SOMs. Next, similar units (i.e., Euclidean distances of the weight vectors are close) among 15 mapping units of SOMs are integrated into the same category using Fuzzy ART. By sorting the facial expression categories integrated by Fuzzy ART from neutral facial expression to the maximum of facial expression, we obtain ELs labeled as expressive intensities of facial expressions quantitatively. The integrated category sorting procedure is based on the two-dimensional correlation coefficient of the average image of the facial expression images classified into each category. Finally, we conduct correspondence of ELs with each frame of the facial images to assess a time-series dataset of variation of ELs.

B. Quantification of Exposed Rhythms using Mutual Information

Mutual information [22] [23] can express changes between signals with the entanglement and synchrony. It can be regarded as an amount that represents linear and nonlinear dependence between the two time-series datasets. Moreover, it represents information flows and dynamically coupled rings between two signals. Mutual information between these two signals is zero if the two systems for observation target differ completely from independent ones. Applying this scheme to the facial expression process, it is possible to quantify the synchronicity and functional connectivity between facial parts. Figure 2 presents one example of time-series changes of ELs in the "Whole face," "Upper part of face," and "Lower part of face" obtained in Section 4.A. In this study, three ROIs listed below are calculated as the mutual information among facial parts in the expressive process. The time-series changes of ELs with respect to the "Whole face," "Upper part of face," and "Lower part of face" respectively represent $R_w = R_w(t)$, $R_u = R_u(t)$, and $R_d = R_d(t)$. Then, mutual information of each ROI is obtained as described below.

Mutual information between the "Whole face" and "Upper part of face" is $I(R_w; R_u)$:

$$I(R_w; R_u) = H(R_w) + H(R_u) - H(R_w, R_u) \qquad (1)$$

Mutual information between the "Whole face" and "Lower part of face" is $I(R_w; R_d)$:

$$I(R_w; R_d) = H(R_w) + H(R_d) - H(R_w, R_d) \qquad (2)$$

Mutual information between the "Upper part of face" and "Lower part of face" is $I(R_u; R_d)$:

$$I(R_u; R_d) = H(R_u) + H(R_d) - H(R_u, R_d) \qquad (3)$$

In that equation, $H(R_w)$, $H(R_u)$, and $H(R_d)$ respectively represent the entropy of $R_w(t)$, $R_u(t)$, and $R_d(t)$. $H(R_w, R_u)$, $H(R_w, R_d)$, and $H(R_u, R_d)$ respectively denote the joint entropy of both.

V. DATASETS

For this study, we constructed an original and long-term dataset for the specific facial expressions of participants. Details of the experimental protocols are the following. One experiment comprises three steps: step 1 is conducted under a normal state; step 2 is done during viewing of a pleasant video; and step 3 is done during viewing of an unpleasant video. We gave participants the task of watching emotion-evoking videos, causing a pleasant–unpleasant state, and took stress measurements by salivary amylase tests to assess the stress state transiently. In addition, the watching time is about 3 min for each emotion-evoking video. We prepared unpleasant videos (i.e., implant surgery and cruel videos) and pleasant videos (i.e., comedy videos of three types). The subjective assessment of five stages was also conducted at watching videos. For all participants, we fully explained the experiment contents in advance, based on the research ethics policy of our university, and also obtained the consent of experiment participants in voluntary writing of participants. Moreover, from each, we received agreement to publish facial images as part of their experimental participation.

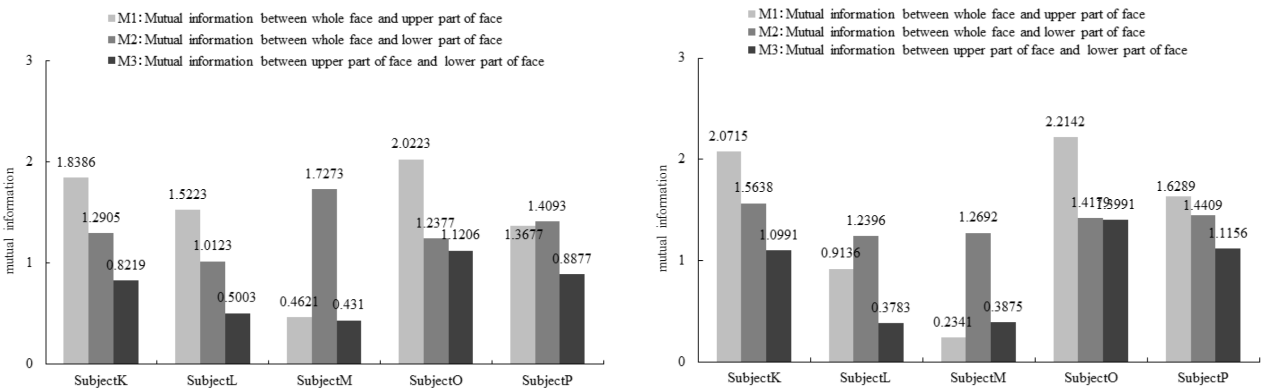
A. Facial Expression Images

Open datasets of facial expression images are open to the public through the internet from universities and research institutes. However, the specifications vary among datasets because of imaging with various conditions. As static facial images, the dataset presented by Ekman and Friesen [24] is a popular dataset comprising collected various facial expressions used for visual stimulation in psychological examinations of facial expression cognition. As dynamic facial images, the Cohn–Kanade dataset [25] and Ekman–Hager dataset [26] are used widely, especially in experimental applications. In recent years, the MMI Facial Expression Database presented by Pantic et al. [27] and the CK+ dataset [28] have become a widely used open dataset containing both static and dynamic facial images. These datasets contain a sufficient number of people as horizontal datasets. However, facial images are taken only once for each person. No dataset exists in which the same person has been traced over a long term. Therefore, we created original and longitudinal datasets that include collections of the specific facial expressions of the same person during a long term.

The six basic facial expressions proposed by Ekman et al. [24] are "happiness," "anger," "sadness," "disgust," "fear," and "surprise." Among those six basic facial expressions, we specifically examined the facial expression of "happiness," which is believed to be most likely to be exhibited spontaneously. As the target facial expression of "happiness" under pleasant and unpleasant stimulation states, we acquired the facial expressions of 20 people. As a stimulation method, we pre-selected emotion-evoking videos that elicit pleasant or unpleasant emotions, with all participants expressing facial expressions of "happiness" immediately after viewing them. Participants, all of whom were university students, were 10 men, whom we designated as A–J (J was 20 years old; B, G, H, and I were 21; A, E, and F were 22; C and D were 23) and 10 women whom we designated as K–T (K, M, O, and P were 20 years old; L, Q, R, S, and T were 21; N was 23). The imaging period was three weeks at one-week intervals for all participants. The imaging environment for facial expressions was an imaging space partitioned by a curtain in the corner of the room. We took frontal facial images with conditions including the head of the participant in each image. In advance, we instructed each participant to expose the facial expression with no head movement. Consequently, imaging the face region to fit within the scope was possible. However, with respect to extremely small changes caused by body motion, we used template-matching methods to trace the face region by setting the initial template to include facial parts. By consideration of the application deployment and ease of imaging in future studies, we used commercially available USB cameras (QcamOrbit; Logicool Inc. [29]). When taking images of each facial expression, the same expression was repeated three times based on the neutral facial expression during the image-taking period of 20 s. We had previously instructed all participants to express an emotion three times at their own timing according to a guideline for 20 s. One dataset consisted of 200 frames with the sampling rate of 10 frames per second.

B. Stress Measurement Method

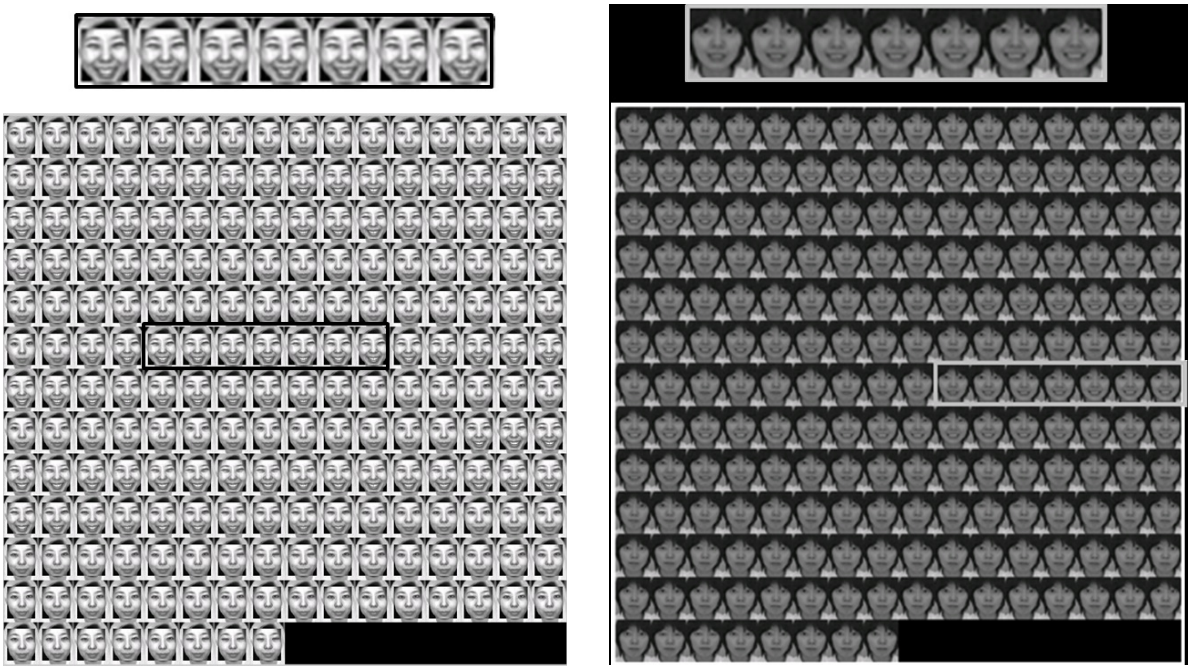
Because types of psychological stress are regarded as affecting facial expressions, we assessed transient stress and chronic stress. Chronic stress is that which humans have on a daily basis, whereas transient stress is that caused by a temporary stimulus. To assess transient stress stimulus to the participants in this study, we applied the salivary amylase test, which measures transient stress reactions. As a biological reaction, salivary amylase activity is detected as a low value if one is in a pleasant state. In contrast, the value is high if one is in an unpleasant state. As stress reactions when subjected to external transient stimulus, Yamaguchi et al. [30] confirmed that salivary amylase activity is an effective means of stress evaluation. For this study, using emotion-evoking videos as an external transient stimulus, we used the salivary amylase test method to measure stress reactions immediately after participants watched the videos.



(a) Pleasant stimulus with emotion-evoking videos

(b) Unpleasant stimulus with emotion-evoking videos

Figure 3. Mutual information results among each facial part for female.



(a) Subject K

(b) Subject M

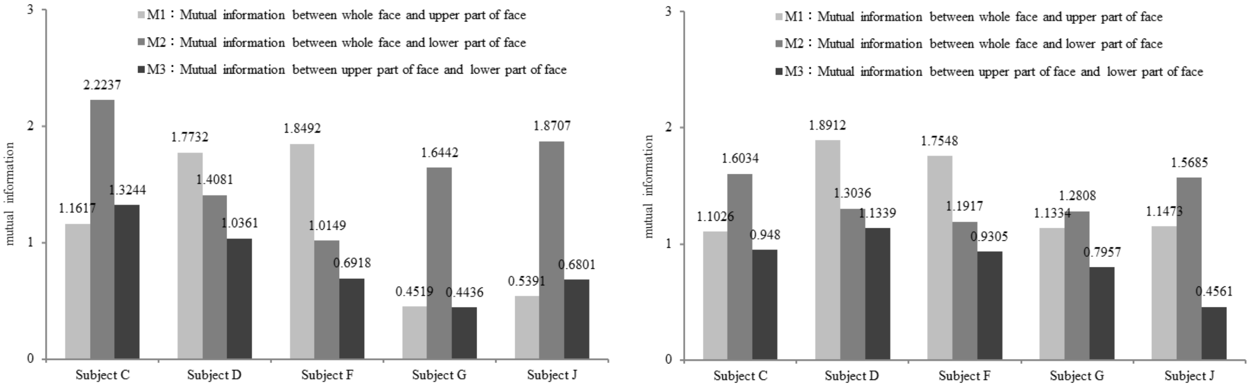
Figure 4. Time-series changes of smile facial expression with pleasant stimulus for specified subjects of female.

VI. EXPERIMENT

Based on the calculation result of mutual information in a time-series change of ELs for each facial region, we analyzed the respective male and female trends. Finally, we discussed the effects of a pleasant–unpleasant stimulus which would give the expressive rhythm of facial expressions from the perspective of mutual information.

A. *Analysis of Female Participants*

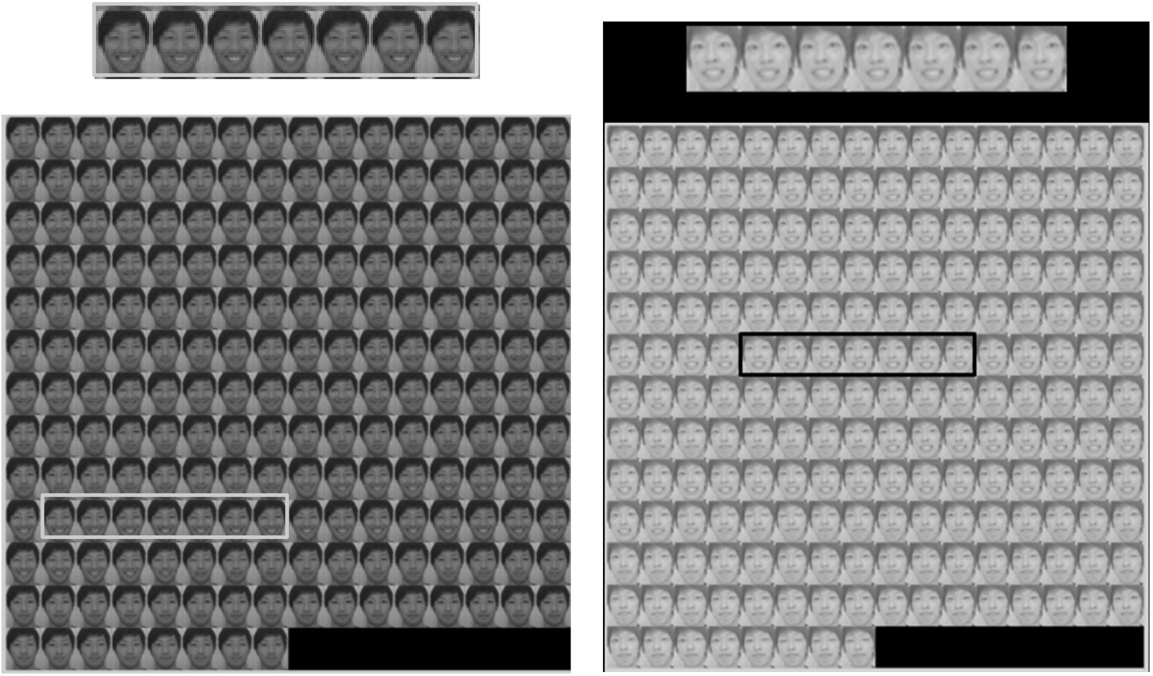
Figure 3 depicts the calculation results of mutual information of five cases of female participants K, L, M, O, and P. The results show the mutual information of the time-series variation of ELs in each face region described in Section 4.B. Figure 3-(a) presents the calculation results obtained after giving a pleasant stimulus. Figure 3-(b) shows calculation results obtained after giving unpleasant stimulus. As an overall trend of female participants, we confirmed the following. For K, L, O, and P, the value of the mutual information is reduced to the order of "ROI 1: between the whole face and upper face," "ROI 2: between the whole face and lower face," and "ROI 3: between the upper face and lower face." The value of "ROI 2: between the whole face and lower face" is clearly larger than those for other ROIs in M. For K, M, and O, we were unable to recognize a marked change in the trend of mutual information by pleasant–unpleasant stimulus. However, for



(a) Pleasant stimulus with emotion-evoking videos

(b) Unpleasant stimulus with emotion-evoking videos

Figure 5. Mutual information results among each facial part for male.



(a) Subject D

(b) Subject J

Figure 6. Time-series changes of smile facial expression with pleasant stimulus for specified subjects of male.

L and P, we detected a specific change in the trend of the mutual information after giving pleasant–unpleasant stimulus. Particularly, the tendency of L is remarkable. In pleasant stimulus, the value of the mutual information is reduced to the order of "ROI 1: between the whole face and upper face," "ROI 2: between the whole face and lower face," and "ROI 3: between the upper face and lower face." Otherwise, "ROI 2: between the whole face and lower face" shows a large value for the unpleasant stimulus. In the unpleasant stimulus, the value of the mutual information is reduced to the order of "ROI 1: between the whole face and upper face," "ROI 2: between the whole face and lower face," and "ROI 3: between the upper face and lower face." However, in pleasant stimulus, the order relation of mutual information of P is reversed with L because the value of "ROI 1: between the whole face and upper face" is reduced.

Next, although the same trend is apparent for both pleasant and unpleasant stimuli, we compare K to M, for which the order relation of the mutual information in each facial region is markedly different. For K in both pleasant and unpleasant stimuli, the mutual information value of "ROI 1: between the whole face and upper face" is larger than "ROI 2: between the whole face and lower face." In addition, particularly addressing "ROI 3: between the upper face and lower face," the value of K is larger than M. However, for M with both pleasant and unpleasant stimuli, the mutual information value of "ROI 2: between the whole face and lower face" is markedly larger than others. Furthermore, particularly addressing "ROI 1: between the whole face and upper face" and "ROI 3: between the upper face and lower face," the values of M are clearly smaller than those of K. For K and M, thumbnail images representing the time-series changes of "happiness" in pleasant stimulus are shown in Figure 4. Figures 4-(a) and 4-(b), respectively present thumbnail images of K and M. The top of each figure shows the characteristic section during exposed facial expression of "happiness." Comparing the thumbnail images shown in Figure 4 to the calculation result of mutual information shown in Figure 3, for K exposed "happiness," we can recognize the change of facial expression in the upper face such as the brow and the area around the eyes, and in the lower face such as the mouth. Otherwise, for M, we can not observe any change of facial expression in the upper face. However, only the corner of mouth in the lower face has changed significantly. Actually, K has the characteristics which the upper part and lower face change both synchronized during facial expressions. Staying on the subjective impression of the experimenter, the result for "happiness" looks more natural facial expressions. In contrast, for M, only the corner of the mouth in the lower face has been changed. Therefore, we have an uncomfortable feeling about the unnatural facial expression of "happiness."

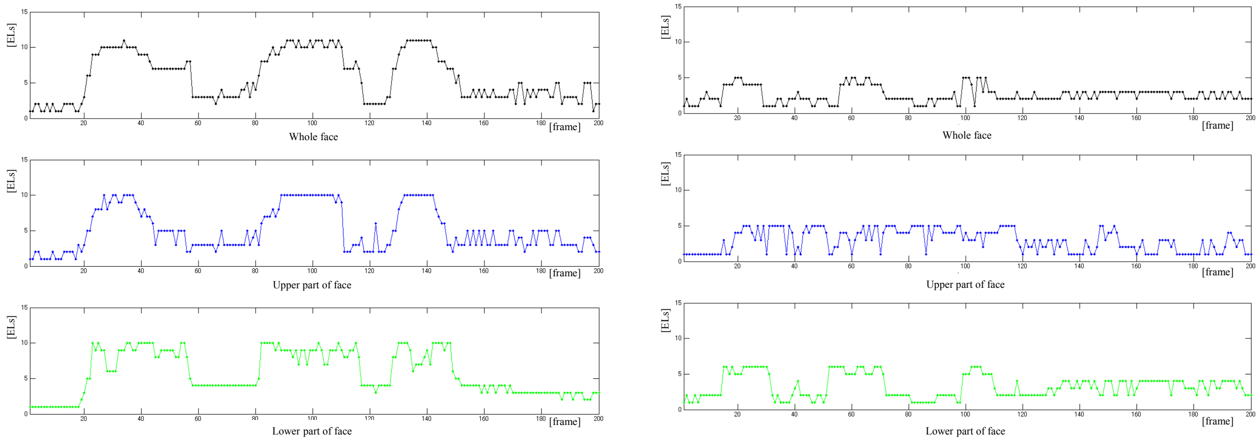
B. *Analysis of Male Participants*

Figure 5 presents calculation results of mutual information of five cases of male participants. Figure 5-(a) presents the calculation results after giving pleasant stimulus. Figure 5-(b) shows the calculation results after giving unpleasant stimulus. As an overall trend of male participants, we confirmed the following. For D and F, the value of the mutual information is reduced to the order of "ROI 1: between the whole face and upper face," "ROI 2: between the whole face and lower face," and "ROI 3: between the upper face and lower face." The value of "ROI 2: between the whole face and lower face" is markedly larger than those of other ROIs in C, G, and J. For male participants C, D, F, G and J, we were unable to recognize a marked change in the trend of mutual information by pleasant–unpleasant stimulus.

Next, regarding male participants, we compare D to J, for whom the order relation of the mutual information in each facial region is significantly different. For D in both pleasant and unpleasant stimuli, the mutual information value of "ROI 1: between the whole face and upper face" is larger than "ROI 2: between the whole face and lower face." In addition, particularly addressing "ROI 3: between the upper face and lower face," the value of D is larger than that of J. However, for J in both pleasant and unpleasant stimulus, the mutual information value of "ROI 2: between the whole face and lower face" is markedly larger than others. In addition, particularly addressing "ROI 1: between the whole face and upper face" and "ROI 3: between the upper face and lower face," the values of J are clearly smaller than those of D. For D and J, the thumbnail images representing the time-series changes of "happiness" in pleasant stimulus are portrayed in Figure 6. Figures 6-(a) and 6-(b) respectively present thumbnail images of D and J. The top of each figure shows the characteristic section during exposed facial expression of "happiness." Comparing the thumbnail images shown in Figure 6 to the calculation result of mutual information shown in Figure 5, for D exposed "happiness," we can recognize the change of facial expression in the upper face such as the brow and around the eyes, and in the lower face such as mouth. Otherwise, for J, no change of facial expression can be observed in the upper face. Only the corner of the mouth in the lower face has changed substantially. Actually, D has characteristics by which the upper part and lower face change at the same time during facial expressions. Therefore, the exposing result of "happiness" looks more natural facial expressions. In contrast, for J, only the corner of the mouth in the lower face has been changed. Therefore, we have an uncomfortable feeling about the unnatural facial expression of "happiness." These results underscore a common tendency between male and female participants and can be anticipated as a new index for quantification of the impression during facial expressions based on the mutual information of the time-series change of each face region.

C. *Effects of Pleasant–unpleasant Stimulus on Mutual Information*

The discrepancy expression in facial expressions means to expose the emotions that do not match one’s own feelings when experiencing certain emotions, such as having a smile, even though one might be in a sad mood. In previous studies, being positive emotional expressions during negative emotional experiences has been shown to engender the following: an amplification of actor’s sympathetic nerve activities [31], an increase of subjective emotional experiences, and some memory loss [32]. The discrepancy expression can easily take cognitive loads for expressive person. Additionally, it can potentially give bad effects to the mental health of actors. Furthermore, the expressive suppression in facial expressions indicates an emotional suppression by facial expressions when experiencing a certain emotion, such as to stifle crying when in a sad mood. Expressive suppression is reportedly associated with social support, closeness with others, and reduction in social satisfaction [33]. In comparison to men, women are more skilled at making smiles and excellent adjustments of positive emotional expressions. Moreover, women show similar effects such as natural expressions to recipients [3].



(a) Subject K (b) Subject M

Figure 7. Comparison of time-series changes of ELs with unpleasant stimulus.

Exposing facial expressions related to "happiness" after viewing an unpleasant video is equivalent to a discrepancy expression. In contrast, exposing the facial expression of "happiness" after viewing a pleasant video is a matching expression. For female participants K and M, such order of mutual information was markedly different; Figure 7 presents their facial expression rhythms. In the impression analysis of Section 6.B, the smile of K gave us a natural impression. In contrast, we received an unnatural impression from the smile of M. Focusing on an expressive rhythm of each facial part, the expressive rhythm of K indicates a time-series change such as to work together in each facial part. In contrast, we were unable to recognize cooperative movements at all in the expressive rhythm of M because the upper face and the lower face are independent. The mutual information of the ROIs (i.e., ROI 1, ROI 2, and ROI 3) effectively expresses the degree of similarity and synchronization of signal waveforms in facial expression rhythms. These mutual information values can be interpreted as quantified indices of the timing structure indicating the synchronization between the upper face (e.g., the eyebrows and eyes) and the lower face (e.g., mouth), which contribute the impression formation to the whole face. We should comprehensively consider the analysis results of Sections 6.B and 6.C. By particularly addressing the magnitude relation between ROI 1 and ROI 2 with respect to the mutual information, we were able to interpret "Eyes say things sufficient to mouth" quantitatively. Around the value of ROI 3 quantifying the timing structure between the upper face and the lower face, noting the magnitude relation and order relation between the values of ROI 1 and ROI 2, it is effective as an index for quantifying the degree of spontaneity and artificiality in facial expressions. Furthermore, more than male participants, the female participants easily created facial expressions of "happiness" intentionally. Then we assumed that result was only slightly affected by the discrepancy expression.

VII. Conclusion and Future Work

In this study, to acquire image datasets of facial expressions under the states of pleasant–unpleasant stimulus for 20 participants (i.e., 10 men, 10 women), we used salivary amylase tests to validate emotional factors when viewing emotion-evoking videos. Additionally, by quantitative analysis of expressive rhythms from the viewpoint of mutual information, particularly addressing expressive processes of "happiness" facial expression after giving a pleasant–unpleasant stress stimulus by emotion-evoking videos, we objectively strove to ascertain complexity and ambiguity when making facial expressions because of human psychological states. Using evaluation experiments examining 10 participants (i.e., 5 men, 5 women), we analyzed the information of time-series changes in ROIs (i.e., ROI 1, ROI 2, and ROI 3), revealing the following points. By particularly addressing the expressive rhythm of each face region, one can estimate the impression of facial expressions from the magnitude relation and order relation of mutual information of each ROI. Additionally, the mutual information of expressive rhythms is effective as an index for measuring degree of spontaneity and artificiality during facial expressions. Female participants were better able to create facial expressions of "happiness" easily and intentionally than male participants were. Moreover, they were less susceptible to discrepancy expressions. In future work, by quantifying fluctuations of expressive tempos in facial parts upon the impression formation, and analyzing their timing structure, we intend to clarify differences of expressive paths between intentional and spontaneous facial expressions.

Acknowledgment

The authors thank the 20 students at our university who participated by letting us take facial images over such a long period. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 25330325 and the Cosmetology Research Foundation.

References

[1] T. Iguchi, “Geometrical Features of the Attractive Smile: –Attractive Production by Kansei X Technology = Kanseiweab–,” The Institute of Electronics, Information, and Communication Engineers, Technical Report, Mar. 2007, pp. 51–56.

[2] T. Yamada and I. Sasayama, “A Study of the Correlation between the Impression Formed from Each Features and the Impression Formed from Face,” Bulletin of Fukuoka University of Education, Vol. 48, No. 4, 1999, pp. 229–239.

[3] L. Ellis, “Gender differences in smiling: An evolutionary neuroandrogenic theory,” Physiology and Behavior, Vol. 88, 2006, pp. 303–308.

[4] K. M. Prkachin, “Effects of deliberate control on verbal and facial expressions of pain,” Pain, Vol. 114, 2005, pp. 328–338.

[5] H. Madokoro, K. Sato, and S. Kadowaki, “Facial expression spatial charts for representing time-series changes of facial expressions,” Japan Society for Fuzzy Theory, Vol. 23, No. 2, 2011, pp. 157–169.

[6] H. Madokoro and K. Sato, “Facial Expression Spatial Charts for Representing of Dynamic Diversity of Facial Expressions,” Journal of Multimedia, Vol. 6, No. 1, Jan. 2007, pp. 1–12.

[7] K. Sato, H. Madokoro, and S. Kadowaki, “Transient Stress Stimulus Effects on Intentional Facial Expressions,” Japan Society for Fuzzy Theory, RJ-005, 2012, pp. 29–36.

[8] K. Sato, H. Otsu, H. Madokoro, and S. Kadowaki, “Analysis of Psychological Stress Factors on Intentional Facial Expressions,” Japan Society for Fuzzy Theory, RJ-002, 2013, pp. 21–28.

[9] K. Sato, H. Otsu, H. Madokoro, and S. Kadowaki, “Analysis of Psychological Stress Factors and Facial Parts Effect on Intentional Facial Expressions,” Proceedings of the Third International Conference on Ambient Computing, Applications, Services and Technologies , Oct. 2013, pp. 7–16.

[10] K. Sato, H. Otsu, H. Madokoro, and S. Kadowaki, “Analysis of Psychological Stress Factors by Using Bayesian Network,” Proceedings of 2013 IEEE International Conference on Mechatronics and Automation, Aug. 2013, pp. 811–818.

[11] H. Ishi, M. Kamachi, and J. Gyoba, “Effect of Facial Motion on Impression of Smile,” The Institute of Electronics, Information, and Communication Engineers, Technical Report, Dec. 2004, pp. 25–30.

[12] S. Hanibuchi, K. Ito, and S. Nishida, “Analysis of Transformed Impression of Smile Process: – An Approach to Supporting Facial Expression Process Training –,” The Institute of Electronics, Information, and Communication Engineers, Technical Report, Oct. 2009, pp. 35–40.

[13] H. Fujishiro, A. Maejima, and S. Morishima, “Natural Smile Synthesis Considering Impression of Facial Expression Process,” The Institute of Electronics, Information, and Communication Engineers, Technical Report, Mar. 2011, pp. 31–36.

[14] H. Fujishiro, A. Maejima, and S. Morishima, “Analysis of Relation between Movement of Smile Expression Process and Impression,” The Journal of the Institute of Electronics, Information, and Communication Engineers, Vol. J95-A, No. 1, 2012, pp. 128–135.

[15] K. W. Kampe, C. D. Frith, R. J. Dolan, and U. Frith, “Reward value of attractiveness and gaze,” Nature, Vol. 413, Oct. 2001.

[16] Y. Kuroki, S. Shiraishi, N. Mukawa, M. Yuasa, and N. Fukayama, “Interaction between Human and Human-like Agent with Gaze and Facial Expression for Human Computer Interaction,” The Institute of Electronics, Information, and Communication Engineers, Technical Report, Mar. 2005, pp. 49–54.

[17] Y. Kuroki, S. Shiraishi, N. Mukawa, M. Yuasa, and N. Fukayama, “Impression of Human-like Agent with Gaze and Facial Expression: –Brain Activity Analysis of HCI using fMRI–,” The Institute of Electronics, Information, and Communication Engineers, Technical Report, Mar. 2006, pp. 43–48.

[18] J.A. Russell and M. Bullock, “Multidimensional Scaling of Emotional Facial Expressions: Similarity from Preschoolers to Adults,” Journal of Personality and Social Psychology, Vol. 48, 1985, pp. 1290–1298.

[19] T. Kohonen, Self-organizing maps, Springer Series in Information Sciences, 1995.

[20] G. A. Carpenter, S. Grossberg, and D.B. Rosen, “Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system,” Neural Networks, vol. 4, 1991, pp. 759–771.

[21] M. Haghighat, S. Zonouz, and M. Abdel-Mottaleb, “Identification Using Encrypted Biometrics,” Computer Analysis of Images and Patterns, Springer Berlin Heidelberg, 2013, pp. 440–448.

[22] T. Ikeda, H. Ishiguro, and M. Asada, “Moving Signal-Source Tracking Based on Mutual Information Maximization,” The Transactions of the Institute of Electronics, Information, and Communication Engineers D, Vol. J90-D, No. 2, 2007, pp. 535–543.

[23] T. Kikuchi, K. Kishi, and J. Miyamichi, “An Automatic Data Classification Algorithm Adjusted by Mutual Information, ” The Transactions of the Institute of Electronics, Information, and Communication Engineers D, Vol. J82-D, No. 4, 1999, pp. 660–668.

[24] P. Ekman and W. V. Friesen, “Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues,” Malor Books, 2003.

[25] T. Kanade, J. F. Cohn, and Y. L. Tian, “Comprehensive database for facial expression analysis,” Proc. of the Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2000, pp. 46–53.

[26] M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, “Measuring facial expressions by computer image analysis,” Psychophysiology, Vol. 36, 1999, pp. 253–264.

[27] M. Pantic, M.F. Valstar, R. Rademaker, and L. Maat, “Web-based Database for Facial Expression Analysis,” Proc. IEEE Int'l. Conf. Multimedia and Expo, Amsterdam, The Netherlands, Jul. 2005. doi: 10.1109/ICME.2005.15214.

[28] P. Lucey et al., “The Extended Cohn–Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression,” Proc. of the Third Int. Workshop on CVPR for Human Communicative Behavior Analysis, 2010, pp. 94–101.

[29] QcamOrbit; Logicool Inc., http://www.logicool.co.jp/ja-jp/webcam-communications/webcams [retrieved: July, 2015]

[30] M. Yamaguchi, T. Kanamori, M. Kanemaru, Y. Mizuno, and H. Yoshida, “Correlation of Stress and Salivary Amylase Activity,” Japanese Journal of Medical Electronics and Biological Engineering: JJME, Vol. 39, No. 3, Sep. 2001, pp. 46–51.

[31] J. L. Robinson and H. A. Demaree, “Physiological and cognitive effects of expressive dissonance,” Brain and Cognition, Vol. 63, 2007, pp. 70–78.

[32] W. Sato, M. Noguchi, and S. Yoshikawa, “Emotion elicitation effect of films in a Japanese sample,” Social Behavior and Personality, Vol. 35, 2007, pp. 863–874.

[33] S. Srivastava, M. Tamir, K. M. McGonigal, O. P. John, and J. J. Gross, “The social costs of emotional suppression: A prospective study of the transition to college,” Journal of Personality and Social Psychology, Vol. 96, 2009, pp. 883–897.

A Color Constancy Model for Non-uniform Illumination based on Correlation matrix

Takashi Toriu
Graduate School of Engineering
Osaka City University
Osaka, Japan
e-mail: toriu@
info.eng.osaka-cu.ac.jp

Mikiya Hironaga
School of Science and Engineering
Kinki University
Osaka, Japan
e-mail: mhironaga@
info.kindai.ac.jp

Hiroshi Kamada
Graduate School of Engineering
Kanazawa Institute of Technology
Ishikawa, Japan
e-mail: kamada@
neptune.kanazawa-it.ac.jp

Thi Thi Zin
Faculty of Engineering
University of Miyazaki
Miyazaki, Japan
e-mail: thithi@
cc.miyazaki-u.ac.jp

***Abstract*—In this paper, we propose a novel color constancy model that works well even if illumination is not uniformly distributed. For this purpose, we introduce the positionally modified color correlation matrix. The color correlation matrix is a matrix that represents how different colors are correlated with one another. More concretely, the color correlation matrix is obtained as the spatial average of the product of two colors as <I_iI_j>, and consequently, it is independent of position parameters. In addition to this correlation matrix, we define two position dependent correlation matrices as <xI_iI_j> and <yI_iI_j>. We assume that the eigenvector of the correlation matrix corresponding to the largest eigenvalue presents the color gray when the illumination is parallel and white. Under this assumption, we can estimate the image when the illumination is parallel and white. The effectiveness of the proposed method is confirmed by simulation experiments using synthesized images and real images.**

Keywords-Color constancy; correlation matrix; non-uniform illumination; gray-world assumption; correlation between brightness and color.

I. INTRODUCTION

Color constancy is one of miracle abilities in human vision. Object colors are correctly perceived independent of the illumination color. This ability is called color constancy [1]. In the field of computer vision, color recognition is an important and basic task. In fact, color recognition has been used as preprocessing for various problems such as robot vision, object recognition, human behavior recognition, human interface and so on. For example, Kamada et al. [2] proposed a system that can count students' raising the color cards to accelerate communication between a teacher and many students in a classroom. In this system, it is very important to achieve accurate color recognition.

Several methods have been proposed for color constancy. Among these methods, the ones based on Gray-world assumption have been popular and they are considered the basis of arguing color constancy [3]-[6]. Gray-world assumption states that the average of the colors of the objects in the scene is gray, and the influence of the illumination is eliminated based on this assumption. This method would work well when sufficient colors exist in the scene. As an extension, a method based on local averaging was proposed by Gijsenij et al. [7]. Further, other methods have been proposed using image statistics such as correlation between the brightness and color [8]-[12]. Golz et al. [8] discussed human color constancy based on the correlation between luminance and redness and concluded that redness of the illumination could be correctly estimated using the mean redness of the image and the correlation between luminance and redness of the image. Inspired by the paper [8], we developed a computational method to estimate illumination color by replacing human eyes with a camera [11]. Previously [13], we proposed another method based on “Minimum Brightness Variance Assumption”, in which it is assumed that variance of brightness is minimum when the illumination is white.

In another previous work [14], we proposed a color constancy model based on the color correlation matrix. More concretely, we proposed two methods for color constancy. Both are based on the correlation matrix on the three-dimensional space of colors, red, green and blue. In the first method, the eigenvector corresponding to the largest eigenvalue is assumed to be a good estimate of the illumination color. In the second method, it is assumed that the eigenvector corresponding to the largest eigenvalue presents the color gray when the illumination is white. The image under white illumination is predicted so as to satisfy the condition that the eigenvector corresponding to the largest eigenvalue presents the color gray.

In these methods, we assumed that illumination is parallel, and therefore it is uniform against position variation. In this paper, we extend the previous method so that it works well for non-uniform illumination. For this purpose, we introduce the positionally modified color correlation matrix. The effectiveness of the proposed method is confirmed by simulation experiments using synthesized images and real images.

ISBN: 978-1-61208-432-9

This paper is organized as follows. In Section II, we outline of the previous method in which illumination is assumed to be parallel and be uniformly distributed spatially. In Section III, we extend this method in the case for non-uniform illumination. In Section IV, we show the experimental results which reveal the effectiveness of the proposed method.

II. Outline of the Previous Method

Let $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ be three unit vectors representing color orientations of red, green and blue in the three-dimensional color space. Then, the unit vector $\mathbf{e}^{(L)} = 1/\sqrt{3}(\mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3)$ represents the orientation of the white color. It is represented as $\mathbf{e}^{(L)} = 1/\sqrt{3}(1,1,1)^t$ in terms of components. Let $I_1(x,y)$, $I_2(x,y)$ and $I_3(x,y)$ be the color components red, green and blue of the input image at the point (x,y), respectively. This image can be represented as

$$\mathbf{I}(x,y) = I_1(x,y)\mathbf{e}_1 + I_2(x,y)\mathbf{e}_2 + I_3(x,y)\mathbf{e}_3\,, \tag{1}$$

in the three dimensional color space. We represent it simply as $\mathbf{I}(x,y) = (I_1(x,y), I_2(x,y), I_3(x,y))^t$. Then, 3 x 3 color correlation matrix K is defined as

$$K_{ij} =< I_i I_j >= \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} I_i(x,y) I_j(x,y) dxdy\,. \tag{2}$$

where $< I_i I_j >$ is the spatial average of $I_i(x,y)I_j(x,y)$. We assume that the image is defined in the rectangle area of $-1/2 \le x, y \le 1/2$.

Let $\mathbf{S}(x,y) = (S_1(x,y), S_2(x,y), S_3(x,y))^t$ be object color and let $\mathbf{E} = (E_1, E_2, E_3)^t$ be illumination color. We assume that illumination is parallel, and therefore it is uniform against position variation. Then, the image $\mathbf{I}(x,y) = (I_1(x,y), I_2(x,y), I_3(x,y))^t$ is determined as

$$I_i(x,y) = E_i S_i(x,y)\,. \tag{3}$$

We call it the image generation formula. From (2) the color correlation matrix K is obtained as

$$K_{ij} =< I_i I_j >= E_i E_j < S_i S_j >\,. \tag{4}$$

Using this relation, the illumination color $\mathbf{E} = (E_1, E_2, E_3)^t$ is determined so that the eigenvector corresponding to the largest eigenvalue of $< S_i S_j >$ is parallel to $\mathbf{e}_l = 1/\sqrt{3}(1 \quad 1 \quad 1)^t$ for the given $K_{ij} =< I_i I_j >$. Once $\mathbf{E}$ is determined, the image under white illumination is estimated as

$$\hat{I}_i(x,y) = cI_i(x,y)/E_i\,, \tag{5}$$

where c is a factor which is determined to keep the brightness invariant. In other words, $\hat{\mathbf{I}}(x,y)$ is considered to be an estimate of the object color $\mathbf{S}(x,y)$. In practice, $\hat{\mathbf{I}}(x,y)$ is obtained as follows.

First, let $\hat{\mathbf{I}}(x,y) = (\hat{I}_1(x,y), \hat{I}_2(x,y), \hat{I}_3(x,y))^t$ be the color components of the image to be estimated and let $\widetilde{K}$ be the correlation matrix of these. $\widetilde{K}$ is obtained as

$$\hat{K} = \frac{1}{N}\sum_{x,y}\hat{\mathbf{I}}(x,y)\hat{\mathbf{I}}^t(x,y)\,. \tag{6}$$

Since we assume that the eigenvector of $\widetilde{K}$ corresponding to the largest eigenvalue is parallel to the orientation of color gray, $\hat{\mathbf{I}}(x,y)$ is estimated so that $\mathbf{e}_l = 1/\sqrt{3}(1 \quad 1 \quad 1)^t$ is the eigenvector of $\hat{K}$ with the largest eigenvalue. More specifically, $\hat{\mathbf{I}}(x,y)$ is obtained by the following procedure. First, as an initialization, we set

$$\hat{\mathbf{I}}(x,y) = \mathbf{I}(x,y)\,, \tag{7}$$

where $\mathbf{I}(x,y)$ is the input image. Second, calculate $\widetilde{K}$ according to the equation (6), and obtain the eigenvector $\mathbf{u}$ of $\widetilde{K}$ with the largest eigenvalue. Then, update $\hat{\mathbf{I}}(x,y)$ according to

$$\hat{I}_i(x,y) \leftarrow \alpha\frac{\beta u_i + 1}{2u_i}\hat{I}_i(x,y)\,, \tag{8}$$

where α is determined to keep the brightness invariant, and β is a factor to modulate the speed of the image change. The larger β is, the smaller the image change is. The above procedure should be terminated when the amount of image change according to (8) is less than a pre-determined value. The amount of image change is evaluated by the root-mean-square error (RMSE).

III. Propsed Method

In the method outlined in the last section, the illumination is assumed to be parallel implicitly. Therefore it is uniform against position variation. In this section, we extend it in the case for non-uniform illumination. More specifically, we assume that position dependence of the illumination is linear and is represented as

$$E_i(x,y) = F_i(1 + \varepsilon_{x_i}x + \varepsilon_{y_i}y)\,. \tag{9}$$

Then, (3) is modified as

$$I_i(x,y) = F_i(1+\varepsilon_{x_i}x+\varepsilon_{y_i}y)S_i(x,y)\,. \tag{10}$$

and correspondingly the correlation matrix becomes

$$\begin{aligned} K_{ij} &=< I_i I_j >= \\ &= F_i F_j < (1+\varepsilon_{x_i}x+\varepsilon_{y_i}y)S_i S_j(1+\varepsilon_{x_j}x+\varepsilon_{y_j}y) > \\ &= F_i F_j < S_i S_j > \\ &\quad + F_i F_j(\varepsilon_{x_i}\varepsilon_{x_j} < x^2 S_i S_j > +\varepsilon_{y_i}\varepsilon_{y_i} < y^2 S_i S_j >) \\ &= F_i F_j(1+\frac{1}{12}\varepsilon_{x_i}\varepsilon_{x_j}+\frac{1}{12}\varepsilon_{y_i}\varepsilon_{y_j}) < S_i S_j >, \end{aligned} \tag{11}$$

where we assume that the image is defined in the rectangle area of $-1/2 \le x, y \le 1/2$ and that

$$\begin{aligned} &< xS_iS_i >= 0, < yS_iS_i >= 0, < xyS_iS_i >= 0, \\ &< x^2S_iS_i >= \frac{1}{12} < S_iS_i >, < y^2S_iS_i >= \frac{1}{12} < S_iS_i >. \end{aligned} \tag{12}$$

Here, we introduce the positionally modified color correlation matrix as

$$\begin{aligned} K_{x_{ij}} &=< xI_iI_i > \\ &= F_iF_j < x(1+\varepsilon_{x_i}x+\varepsilon_{y_i}y)S_iS_j(1+\varepsilon_{x_j}x+\varepsilon_{y_j}y) > \\ &= F_iF_j(\varepsilon_{x_i}+\varepsilon_{x_j}) < x^2S_iS_j > \\ &= \frac{1}{12}F_iF_j(\varepsilon_{x_i}+\varepsilon_{x_j}) < S_iS_j > \end{aligned} \tag{13}$$

and

$$\begin{aligned} K_{y_{ij}} &=< yI_iI_i > \\ &= F_iF_j < y(1+\varepsilon_{x_i}x+\varepsilon_{y_i}y)S_iS_j(1+\varepsilon_{x_j}x+\varepsilon_{y_j}y) > \\ &= F_iF_j(\varepsilon_{y_i}+\varepsilon_{y_j}) < y^2S_iS_j > \\ &= \frac{1}{12}F_iF_j(\varepsilon_{y_i}+\varepsilon_{y_j}) < S_iS_j > \end{aligned} \tag{14}$$

From (11), (13) and (14) we obtain a system of equations for ε_{i_x} and ε_{i_y} as follows.

$$\begin{aligned} \frac{K_{x_{ij}}}{K_{ij}} &= \frac{\varepsilon_{x_i}+\varepsilon_{x_j}}{\varepsilon_{x_i}\varepsilon_{x_j}+\varepsilon_{y_i}\varepsilon_{y_j}+12}, \\ \frac{K_{y_{ij}}}{K_{ij}} &= \frac{\varepsilon_{y_i}+\varepsilon_{y_j}}{\varepsilon_{x_i}\varepsilon_{x_j}+\varepsilon_{y_i}\varepsilon_{y_j}+12}. \end{aligned} \tag{15}$$

This system of equations has 12 equations against 6 unknown parameters ε_{i_x} and ε_{i_y}. Among them we use only $i = j$ case (6 equations) for simplicity. Then, (15) becomes

$$\begin{aligned} T_{x_i} &= \frac{2\varepsilon_{x_i}}{\varepsilon_{x_i}{}^2+\varepsilon_{y_i}{}^2+12}, \\ T_{y_i} &= \frac{2\varepsilon_{y_i}}{\varepsilon_{x_i}{}^2+\varepsilon_{y_i}{}^2+12}, \end{aligned} \tag{16}$$

where $T_{x_i} = K_{x_{ii}} / K_{ii}$ and $T_{y_i} = K_{y_{ii}} / K_{ii}$. This system of six equations can be solved easily and we obtain

$$\begin{aligned} \varepsilon_{x_i} &= \frac{T_{x_i}\left(1-\sqrt{1-12(T_{x_i}{}^2+T_{y_i}{}^2)}\right)}{T_{x_i}{}^2+T_{y_i}{}^2}, \\ \varepsilon_{y_i} &= \frac{T_{y_i}\left(1-\sqrt{1-12(T_{x_i}{}^2+T_{y_i}{}^2)}\right)}{T_{x_i}{}^2+T_{y_i}{}^2}. \end{aligned} \tag{17}$$

As is described in (9), illumination is determined by parameters F_i, ε_{i_x} and ε_{i_y}. The remained problem is to obtain F_i. For this purpose, we notice that (10) can be rewritten as

$$\frac{I_i(x,y)}{1+\varepsilon_{x_i}x+\varepsilon_{y_i}y} = F_iS_i(x,y)\,. \tag{18}$$

If we set

$$\tilde{I}_i(x,y) = \frac{I_i(x,y)}{1+\varepsilon_{x_i}x+\varepsilon_{y_i}y}\,. \tag{19}$$

Equation (18) becomes

$$\tilde{I}_i(x,y) = F_iS_i(x,y)\,. \tag{20}$$

This equation has the same form as (3), which represents the image generation formula for the case when illumination is uniform. In view of this, the constant factor

$\mathbf{F}=(F_1,F_2,F_3)^t$ of the illumination color is determined so that the eigenvector corresponding to the largest eigenvalue of $< S_iS_j >$ is parallel to $\mathbf{e}_l = 1/\sqrt{3}(1 \quad 1 \quad 1)^t$ for the given $\tilde{K}_{ij} =< \tilde{I}_i\tilde{I}_j >$.

The algorithm of the proposed method is summarized as follows.

Step 1: Calculate the correlation matrixes as

$$K_{ij} =< I_iI_i >= \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} I_i(x,y)I_j(x,y)dxdy\,,$$

$$K_{xij} =< xI_iI_i >= \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} xI_i(x,y)I_j(x,y)dxdy$$

and

$$K_{y_{ij}} =< yI_iI_i >= \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} xI_i(x,y)I_j(x,y)dxdy\,.$$

Step 2: Set

$$T_{x_i} = K_{x_{ii}} / K_{ii}$$

and

$$T_{y_i} = K_{y_{ii}} / K_{ii}\,.$$

Step 3: Get

$$\varepsilon_{x_i} = T_{x_i}\left(1-\sqrt{1-12(T_{x_i}{}^2+T_{y_i}{}^2)}\right)\Big/\left(T_{x_i}{}^2+T_{y_i}{}^2\right)$$

and

$$\varepsilon_{y_i} = T_{y_i}\left(1-\sqrt{1-12(T_{x_i}{}^2+T_{y_i}{}^2)}\right)\Big/\left(T_{x_i}{}^2+T_{y_i}{}^2\right).$$

Step 4:Calculate the modified correlation matrix as

$$\tilde{K}_{ij} =< \tilde{I}_i\tilde{I}_i > = \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \frac{I_i(x,y)I_j(x,y)}{(1+\varepsilon_{x_i}x+\varepsilon_{y_i}y)(1+\varepsilon_{x_j}x+\varepsilon_{y_j}y)}dxdy\,.$$

Step 5: Find $\mathbf{F}=(F_1,F_2,F_3)^t$ such that the eigenvector corresponding to the largest eigenvalue of the matrix $F_i{}^{-1}\tilde{K}_{ij}F_j{}^{-1}$ is parallel to $\mathbf{e}_l = 1/\sqrt{3}(1 \quad 1 \quad 1)^t$

Step 6: Obtain the image under white illumination as

$$\hat{I}_i(x,y) = c\tilde{I}_i(x,y)/G_i\,,$$

where c is a factor which is determined to keep the brightness invariant.

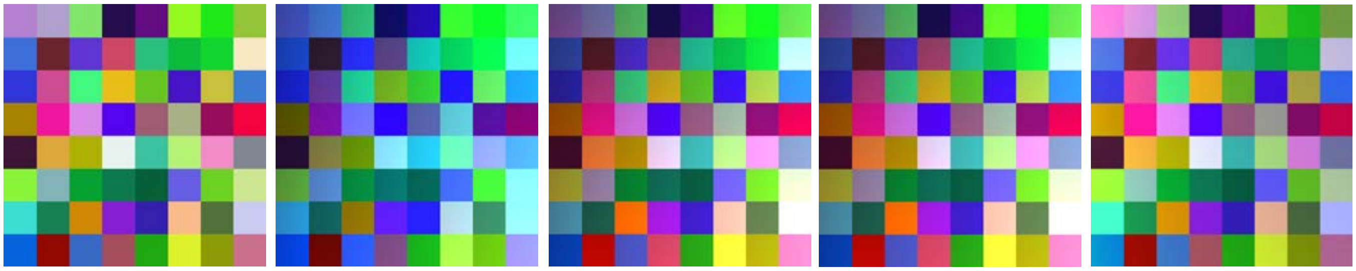

(a) (b) (c) (d) (e)

Figure 1. Examples of images: (a) an object color, (b) an input image, (c) a result by the Gray-world based method, (d) a result by the previous correlation based method and (e) a result by the proposed method.

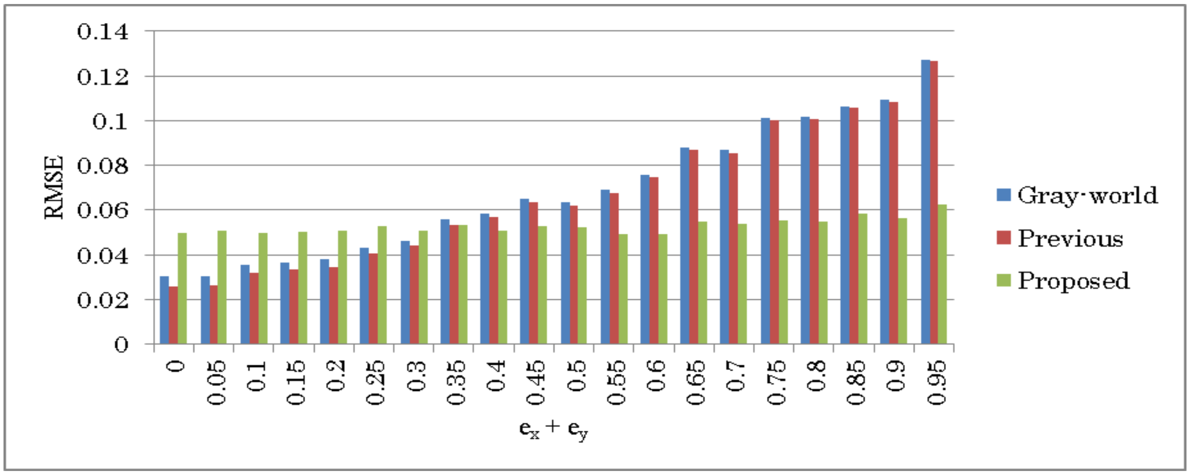


Figure 2. Root-mean-square error between the ideal image and the estimated image in the experiments using synthesized images.

ISBN: 978-1-61208-432-9

IV. EXPERIMENTS

We conducted two types of simulation experiments to confirm the effectiveness of the proposed methods. In the first experiments we synthesized images with 512 × 512 size which represents the object color. The image has 8 × 8 = 64 blocks, and in each block the object colors are specified by $S_1(x,y)$, $S_2(x,y)$ and $S_3(x,y)$, which are determined at random and uniformly distributed from 0.0 to 1.0. In this manner, we prepared 100 samples of images representing object colors for each illumination condition mentioned below.

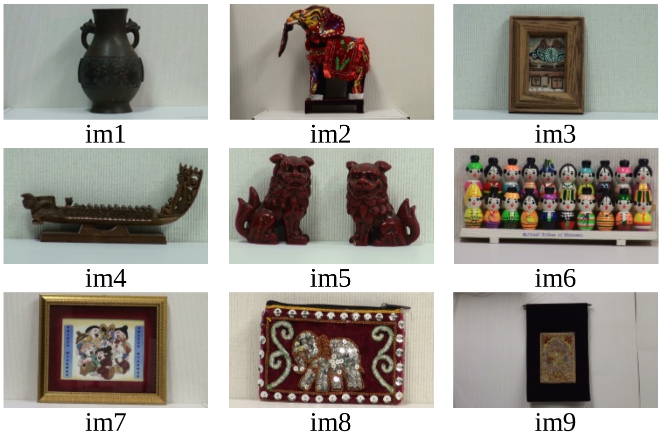


Figure 3.Input images under common room lighting.

An example of synthesized image is shown in Figure 1 (a).Next, we set F_1 , F_2 and F_3 to be 0.4, 0.7, 1.0 respectively as the constant factor of the color component red, green and blue (see equation (9)). Further, we set $\rho_1 = 0.0, \rho_2 = 0.05, \cdots, \rho_{20} = 0.95$, which represent the magnitude of $e_{x_i} + e_{y_i}$. Each e_{x_i} and e_{y_i} is determined randomly as follows. First, e_{x_i} is determined at random and uniformly distributed from 0.0 to ρ_m $(m = 1,2,\cdots 19)$. Then, e_{y_i} is determined as $e_{y_i} = \rho_m - e_{x_i}$. In consideration of origin symmetry, we restricted e_{x_i} and e_{y_i} to be zero or positive without loss of generality. The parameter ρ_m is considered to represent the amount of non-uniformity. Thus 20 kind illumination conditions are defined according to (9). Then, we generate 100 input images for each non-uniformity parameter according to (10) using randomly synthesized 100 sets of object colors $S_1(x,y)$, $S_2(x,y)$ and $S_3(x,y)$. In this way, we got 2000 (100 x 20) input images. An example of input image is shown in Figure 1 (b), where $e_{x_i} + e_{y_i}$ $(i = 1,2,3)$ are set to be 0.95.For these 2,000 images, we estimated images under white illumination based on the proposed method. We also estimated the images using the Gray-world based method and the previously proposed correlation based method for comparison. In Figure 2, the RMSE between the ideal image and the estimated images by the three methods are shown. Examples of estimated images by the three methods are shown in Figure 1 (c), (d) and (e). As can be seen in Figure 2, the proposed method has its own superiority when illumination non-uniformity is large. Inversely, when illumination is uniform, the previous method has rather better results.

Next, we conducted experiments using 9 groups of real images. Each group has 4 images, with one image taken under common room lighting. These are shown in Figure 3. The other three images in each group are taken under such illumination condition in which colored lamps (red, green and blue) are set from the upper right direction in addition to the common illumination. Figure 4 (a) shows an example of the input image under common room lighting. Figure 4 (b), (c) and (d) are examples of images taken under illumination condition where each of red, green and blue lamp is set.

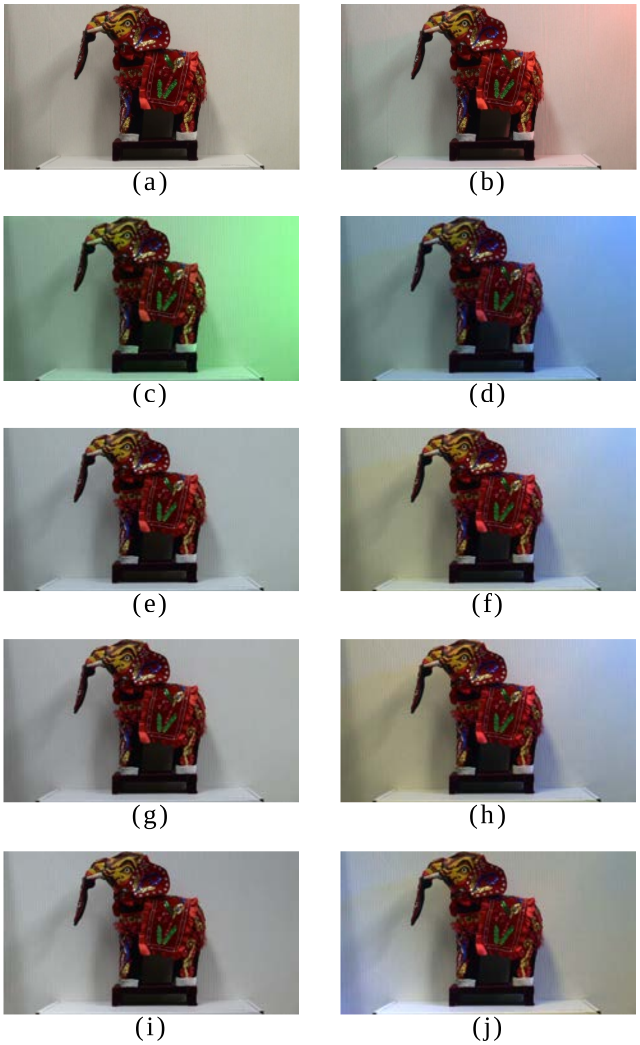


Figure 4. Examples of input images (a) – (d) and estimated images (e)- (j). (a) Common illumination is added. (b) Red illumination is added. (c) Green illumination is added. (d) Blue illumination is added. (e), (g) and (i) show the results for the case when the input image is (a). (f), (h) and (j) show the results in the case when the input image is (d)

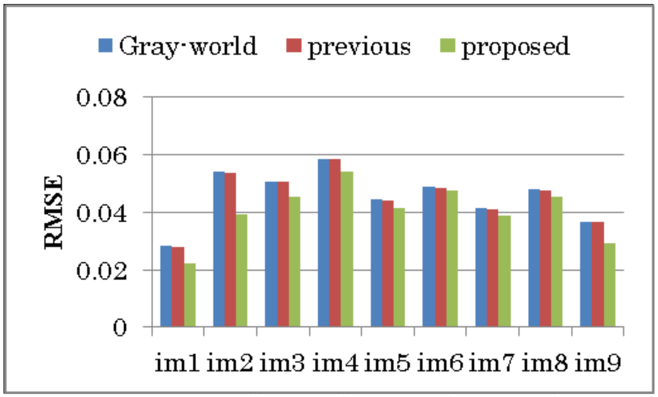


Figure 5. Average RMSE between the estimated image under common illumination and the estimated image under colored illumination.

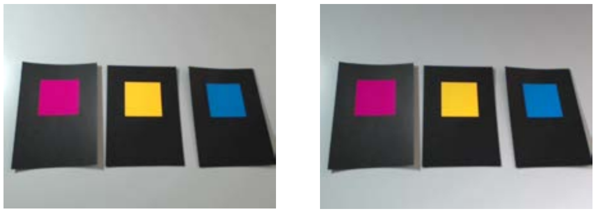


Figure 6. Another example of input images (left hand), and the result by the proposed method (right hand).

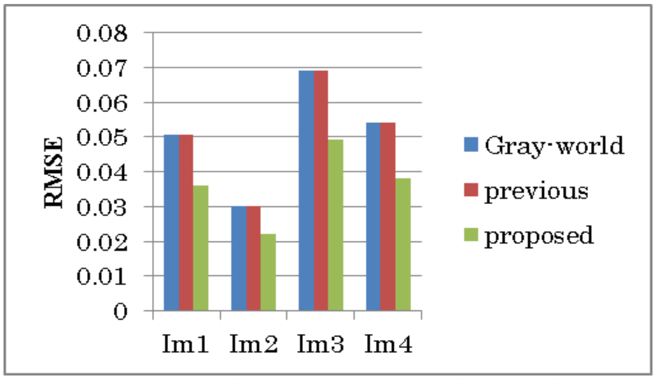


Figure 7. Average RMSE between two the estimated image under common illumination and the estimated image estimated under colored illumination.

Using these images, we conducted experiments to confirm the effectiveness of the proposed method. We compared three methods, Gray-world based method, the previously proposed correlation based method and the method proposed in this paper. To evaluate the effectiveness of each method, we calculate the RMSE between two estimated images. One is the image estimated from the image taken under common illumination. The other is the image estimated from the image taken under colored illumination. RMSEs for three illumination conditions are averaged. The results are summarized in Figure 5. As can be seen, proposed methods have smaller RMSE. It means that the proposed method is effective as a kind of a normalization tool of images to reduce image changes due to illumination changes. This method is not necessarily just a tool to obtain the image under white illumination. More specifically, the amount of change between two estimated images is less than the amount of change between the corresponding two input images with different illumination condition.

Figure 4 (e) – (j) shows examples of the estimated images. Among them (e), (g) and (i) show the results for the case when the input image is (a), the image taken under common room lighting. On the other hand, (f), (h) and (j) show the results in the case when the input image is (d), the image taken under blue illumination. The pairs (e)-(f), (g)-(h) and (i)-(j) correspond to the Gray-world based method, the previous correlation based method and the proposed method. Figure 5 shows average RMSE between two estimated images. One is the image estimated from the image taken under common illumination. The other is the image estimated from the image taken under colored illumination. RMSEs for three illumination conditions are averaged. From this figure, we can see that the RMSE between the image (i) and (j), which is in the proposed method, is about o.o4, which is significantly less than in the Gray-world based method and in the previous correlation based method.

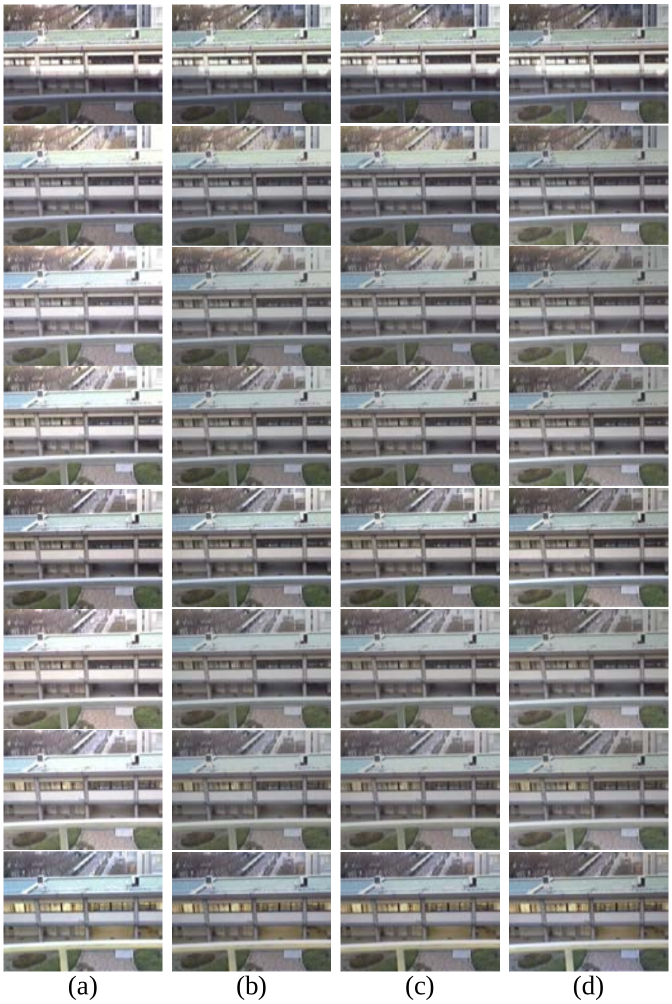


Figure 8 Sample images. (a) Original images. (b) Results by Gray-world based method. (c) Results by the previous method. (d) Results by the proposed method.

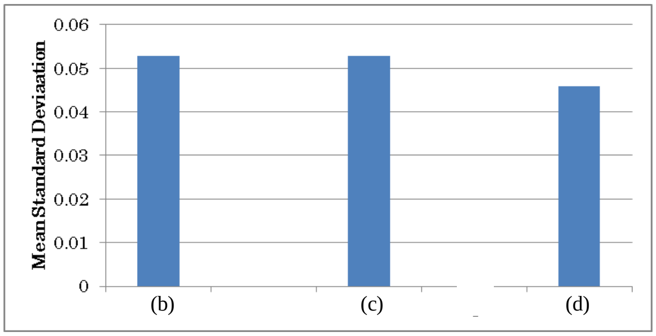


Figure 9 Mean standard deviation of pixel values.

We are planning to apply the proposed method to the system that can count students' raising the color cards to accelerate communication between a teacher and many students in a classroom [2]. An example of the color cards is shown in Figure 6 (a). The upper left part of the image is brighter than other parts. Figure 6 (b) is the result of the proposed method. Non-uniformity of the lighting effect is reduced. We prepared 4 groups of images like Figure 6 (a). Each group has 4 images with different illumination condition. One is taken under a common lighting, and the other three are taken under the common lighting added by another lamp with different strength. We evaluated the effectiveness of the proposed method in the same way as mentioned before, and we obtained the result shown in Figure 7. The proposed method has the lower RMSE. Figure 6 (a) is an image included in the group Im3.

We conducted other experiments to evaluate effectiveness of the proposed method in the case when illumination changes with setting of the sun. Figure 8 (a) shows 8 samples of original images taken in setting of the sun. We can see changes in these images caused by illumination change. These changes were eliminated by three color constancy models; the gray-world based method, the previous method and the proposed method. The results are shown in Figure 8 (a), (b) and (c). We calculated the standard deviation of pixel values over 8 images shown in Figure 8 (a), (b) and (c). Figure 9 shows the mean standard deviation. If this value is small, it means that effect of illumination change is effectively eliminated. From this figure, we can conclude that the proposed method is more effective than other two methods.

V. CONCLUSION

We extended the previous method [14] based on the color correlation matrix $K_{ij} =< I_i I_j >$ so that it works well for non-uniform illumination. In the proposed method, we introduce the modified correlation matrix. $K_{xij} =< xI_i I_j >$ and $K_{xij} =< xI_i I_j >$ to reduce the non-uniformity in the illumination. We conducted simulation experiments using synthesized image and confirmed that the proposed method has its own superiority when non-uniformity of the illumination is large. Inversely, when the illumination is uniform, the previous method has rather better results. We also confirmed that the proposed method is effective for real images too. In future work, we will address the problem that the proposed method does not have better result compared to the previous method when the illumination is uniform.

REFERENCES

[1] D. H .Foster, “Color constancy”, Vision Research, 51, 2011, pp.674-700.

[2] H. Kamada and K. Masuda, “A Feasibility Study of Automatic Response Analyzer in Classroom Using Image Processing and Cards”, ICIC Express Letters, Part B, Vol.6, No.4, 2015, pp.919-926.

[3] M.D'Zmura and P. Lennie “Mechanisms of color constancy”, J. Opt. Soc. Am.A, Vol.3, No.10, Oct. 1986, pp.1662-1672.

[4] K. Barnard, V. Cardei and B. Funt “A comparison of computational color constancy algorithms. Part I: Methodology and experiments with synthesized data”, IEEE Trans. IP, 11, No.9, 2002, pp.972 -984.

[5] H. Kawamura, K. Fukushima and N. Sonehara, “Mathematical conditions for estimating illumination color based on gray world assumption”, Proc. of the Society Conference of IEICE, 167, 1995.

[6] F. Ciurea and B. Funt, “A Large Image Database for Color Constancy Research”, IS&T/SID Eleventh Color Imaging Conference, 2003.

[7] A. Gijsenij and T. Gevers, ”Color Constancy by Local Averaging”, 14th International Conference of Image Analysis and Processing – Workshops, 2007.

[8] M. Golz, ”Influence of scene statistics on color constancy,”Nature, 415, 2002, pp.637–640.

[9] J. Golz., “The role of chromatic scene statistics in color constancy: Spatial integration”, Journal of Vision, 8(13):6, 2008, pp.1–16.

[10] K. Barnard, L. Martin and B. Funt, “Colour by correlation in a three dimensional colour space”, Proc. of the Sixth European Conf. on Comp. Vis., 2000, pp.275–289.

[11] T. Yoshida, M. Hironaga and T.Toriu, “Estimating and Eliminating a Biased Illumination with the Correlation between Luminance and Colors”, ICIC Express Letters, Vol..7, No.5, 2013, pp.1687-1692.

[12] J. J. M. Granzier, E. Brenner, F. W. Cornelissen and J. B. J. Smeets ,“Luminance color correlation is not used to estimate the color of the illumination”, Journal of Vision, 5, 2005, pp.20–27.

[13] N. Hasebe, M. Hironaga and T. Toriu, “A Color Constancy Model with Minimum Brightness Variance Assumption, ICIVC 2014.

[14] T. Toriu, M. Hironaga and N. Hasebe, “Two methods for color constancy based on color correlation matrix”, ICGEC 2015, submitted.

A Semantic Representation for Process-Oriented Knowledge Management to support Production Planning based on Function Block Domain Models and a Three-level Mediator Architecture

Benjamin Gernhardt,
Franz Miltner, Tobias Vogel,
Matthias Hemmje
Multimedia and Internet Applications
University of Hagen
Hagen, Germany
e-mail: firstname.lastname@fernuni-hagen.de

Holger Brocks
InConTec GmbH
Burghaslach, Germany
e-mail: holger.brocks@incontec.de

Lihui Wang
Sustainable Production Systems
Royal Institute of Technology Stockholm
Stockholm, Sweden
e-mail: lihui.wang@iip.kth.se

***Abstract*—Semantic approaches for knowledge representation and management as well as knowledge sharing, access and re-use can support *Collaborative Adaptive Production Process Planning (CAPP)* in a flexible, efficient and effective way. Therefore, semantic-technology based representations of such CAPP knowledge integrated into a machine readable process formalization is a key enabling factor for sharing such knowledge in cloud-based semantic-enabled knowledge repositories supporting CAPP scenarios as required in the CAPP-4-SMEs project. Beyond that, *Small and Medium Enterprises (SMEs)* as represented in CAPP-4-SMEs request for a standardized CAPP-oriented product-knowledge- and production-feature representation. That can be achieved by applying so-called *Function Block (FB)* based knowledge representation models. Web-based and at the same time Cloud-based technologies, tool suites and application solutions which are based on process-oriented semantic knowledge-representation methodologies, such as *Process-oriented Knowledge-based Innovation Management (German: Wissens-basiertes Prozess-orientiertes InnovationsManagement, WPIM)* can satisfy these needs. In this way, WPIM can be applied to support the integration and management, as well as the access and re-use in a machine readable and integrated representation of distributed CAPP knowledge. On the other hand that knowledge is shared within a cloud-based centralized semantic-enabled knowledge repository. Furthermore, semantic knowledge representation and querying will add value to the knowledge-based and computer-aided re-use of such machine-readable knowledge resources within CAPP activities. Finally, it will pave the way towards further automating planning, simulation and optimization in a semantic-web for CAPP.**

Function Blocks; DPP; CAPP; Process Planning; Process-oriented Knowledge Management; Knowledge-based Process-oriented Innovation Management; WPIM

I. INTRODUCTION, MOTIVATION AND PROBLEM STATMENT

In [1], the general concept of developing a knowledge-based and process-oriented CAPP support by using the WPIM method as a basis was proposed. The WPIM approach offers the possibility of modeling and representing innovation processes in a machine-readable semantic format and furthermore enables annotating the process representation in a semantic way with further knowledge resources. This whole representation structure can then later be accessed by means of semantic queries. However, so far WPIM has only been applied in domains like design and development. It also includes ***Product Life Cycle Management (PLM)*** support but it has not yet been practically applied in the domain of CAPP. In parallel to the development of WPIM, Wang et al. have introduced a method for representing web-based ***Distributed Process Planning (DPP)*** activities in [3], [4] and [5]. In the following we will use slightly adapted excerpts from [3] to introduce the necessary concepts and rationale of the DPP method. The DPP method includes also the concepts of ***Meta Function Blocks (MFBs), Execution Function Blocks (EFBs) and Operation Function Blocks (OFBs)***. Furthermore, Helgoson et al. explain in [6] that “Today, machining-feature based approaches combined with artificial-intelligence (AI) based methods are the popular choices for process planners”. Their introduced approach is already based on a DPP modeling-method but does not yet support machine-readability and semantic interoperability of such models as it could be achieved by utilizing representations as available in nowadays semantic web technologies and as e.g., supported by WPIM. This means, while the proposed DPP approach is very useful and valid in terms of representing the product and machining features within MFBs, EFBs and OFBs but nevertheless it does not yet support semantic-web based cross-organizational and cross-domain knowledge sharing. However, this is necessary to make such knowledge more widely available e.g., to be shared in collaborations of SMEs within CAPP activities. This DPP knowledge is not so far available in a machine-readable semantic representation at all. The interoperability of such a representation with technologies of the semantic-web and therefore with other applications and tools, like e.g., from the area of ***Artificial Intelligence (AI)*** and ***Machine Learning (ML)***, cannot easily be achieved.

Moreover, this knowledge cannot easily be automatically shared, managed, accessed, exchanged and re-used within collaborations that take advantage of cloud-based semantic repositories of CAPP-knowledge. In general, therefore can also not easily be accessed online virtually in real-time during computer-supported CAPP activities within latest (SoA) ICT infrastructures across knowledge domains and organizational borders.

At the same time, if such a semantic and process-oriented CAPP-knowledge representation utilizing semantic-web technologies would exist then it could be very well supported by other semantic-web enabled technologies and corresponding methods, like, e.g., WPIM and in this way interoperability. Therefore an integration of cloud-based semantic CAPP knowledge repositories with other e.g., AI and CAPP-support technologies paradigm could be achieved by means of integrating them based on the semantic web software development paradigm.

In consequence, this insight requires the application of semantic technologies and corresponding methods like e.g., WPIM. Process-oriented semantic representation of CAPP knowledge wherein the product and machining features are formalized within MFBs, EFBs and OFBs like domain-specific representations i.e., domain models of the DPP knowledge domain could support the CAPP knowledge domain.

The remainder of this paper is based on this insight and is applying and implementing the necessary DPP and semantic-web integration approach within a mediator architecture. Such architectures are typical for semantic-web repositories and solving semantic integration challenges as well as integrating several local knowledge sources into a global, potentially cloud-based, semantic repository. This can then be considered a semantic and cloud-based CAPP-knowledge repository which has been implemented in a very (technologically) open and distributed way. From the point of view of WPIM, the domain models for MFBs, EFBs and OFBs can be covered by a semantic integration in this repository with the existing WPIM domain concepts of **WPIM-Master Processes**, -**Process Instances**, -**Tasks** and –**Activities** (explained in Section 2.5). Thus, it is allowed for the integration of WPIM- and DPP-based knowledge modeling as well as for the semantic representation of DPP knowledge to become available as a knowledge-based support to CAPP activities. In the rest of this paper, we will also describe this integration more in detail. Because of that, this paper covers the following aspects:

First of all, the state-of-the-art of FB-based production planning models and in detail the proposed DPP method including the necessary planning processes producing and handling MFBs, EFBs and OFBs will be revisited. Furthermore, we describe the state-of-the-art w.r.t. Process Ontologies and more accurate the WPIM-Ontology. We will carry out a comparison of the DPP modeling approach of Wang et al. [3] with the expressiveness of the WPIM-Ontology and we will introduce the prototypical extension of the WPIM-Ontology to cover i.e., semantically wrap and integrate the DPP planning processes and there resources including MFB, EFB and OFB concepts of the DPP-model. We discuss and analyze all these already mentioned different approaches in Section 2. The integrating role of WPIM in the domain of process planning will be explained in Section 3.

Furthermore, we will outline our mediation approach to a DPP-based distributed knowledge representation. This will result in an extension of the WPIM tool suite and application solution by means of a mediator architecture to support the integration into a centralized and potentially cloud-based global CAPP repository. We illustrate this in Sections 4 and 5. In this way, such a repository that can then support in the future cross-domain and cross-organizational CAPP processes, tasks, and activities in terms of knowledge sharing and online process-driven access support.

In Section 6, we present the theoretical basis of our three level mediator architecture. Finally, discussions, future work and conclusions are given in Section 7.

II. State of the Art and Analysis

The following paragraphs will briefly summarize and analyze the state-of-the-art of FB based DPP modeling. The section is based on a slightly adapted excerpt from [3] and WPIM-based semantic process-modeling. It also introduces the necessary concepts of information integration and mediation as well as of mediator architectures as a background for the integration and mediation approach to be applied for the integration of DPP and WPIM.

A. Function Blocks

FBs are initially defined in the IEC 61499 standard [7], which explains the usage, development and implementation of FBs in distributed industrial process measurement and -control systems in a component-oriented approach [8]. IEC 61499 was developed jointly from the existing concepts of FB diagram in the ***Programmable Logic Controllers (PLC)*** language standard IEC 61131-3 [9] and standardization work concerning Fieldbus [9]. It was developed after the need for a common model for the application of software modules called FBs had been raised. FB diagrams were initially introduced (in IEC 61131-3) to solve problems with textual programming, ladder diagrams, and the reuse of common tasks. In the new standard of IEC 61499, an FB is an event-triggered component containing algorithms and an ***Execution Control Chart (ECC)*** with inputs and outputs of data and events. Algorithms are executed when triggered by input events, reading data from the input data and producing new output data. The algorithm execution and scheduling is controlled by the ECC functioning like a finite state machine and at the end of algorithm execution an output event is created. As basic building blocks, many FBs can be combined in a distributed network to create complex control applications with their data/event interfaces interconnected to control the flow of data and events. One FBs output event could then be the input event of another FB. A common way of describing or viewing an in summary, FB can be considered as a model of software or process representation, treating the encapsulated behavior in a form that is similar to an electronic circuit. A literature review related to the FB related research targeting the areas of machining and assembly is available in [3][4], as well as an introduction into Distributed Process

Planning (DPP) as an important stepping stone towards supporting CAPP with DPP methodology.

B. Distributed Process Planning & Meta Function Blocks

Furthermore, as outlined in more detail in [3], the required functionality for implementing a web-based DPP system is consisting of three core components of the DPP, namely the planning processes of ***Supervisory Planning (SP), Operation Planning (OP)*** plus a new ***Execution Control Planning (ECP)***, which are explicitly modeled in a conceptual ***ICAM Definition for Function Modeling (IDEF0***, where '***ICAM***' is an acronym for ***Integrated Computer Aided Manufacturing***) process formalization model together with their inter-relationship and dataflow. Meta Function Blocks (MFBs) are used in this research to encapsulate machining sequences (of setups and machining features) and are the output of Supervisory Planning. As its name suggested, an MFB only contains generic information about process planning of a product. It is a high-level process template, with suggested cutting tool types and tool path patterns, for subsequent manufacturing tasks.

C. Execution and Operation Function Blocks

Within the DPP methodology, Execution Function Blocks (EFBs) are the FBs that are ready to be downloaded to a specific machine. Basically, an EFB can be created by instantiating a series of MFBs associated with a task. Each manufacturing task corresponds to its own set of EFBs, so that the monitoring functions can be conducted for each task unit. Furthermore, the DPP methodology offers the concept of an Operation Function Blocks (OFBs). The structure of an OFB is the same as that of an Execution Function Block (EFB). However, an OFB specifies and completes EFB with more detailed, machine-specific data about machining processes and operation sequences. Moreover, operation planning module can override and update the actual values of variables in the EFB, so as to make it locally optimized and adaptable to various events happened during machining operations. Wang et al. use the two different terms of EFB and OFB in [3] to distinguish a given FB, because they are two separate entities with different level of detail in contents, fulfilling different level of execution, residing in different systems, and moreover, they may be deployed in physically distributed ***Computerized Numerical Control (CNC)*** controllers. In other words, a FB holds a set of predefined algorithms that can be triggered by an arriving event to the FB. Thus, a decision can be made by executing the algorithm.

D. WPIM

The concept of WPIM was developed to support capturing and usage of knowledge around innovation processes [1] [2][10]. It assumes that innovation has both a knowledge and process perspective, which needs to be used in a combined manner. Therefore, activities of a process can be annotated with resources, such as experts and documents [10].

The web-based WPIM application and corresponding tool suite [28] allows the integration and mediation of semantic representations of process structures and specific knowledge resources. To support CAPP, the so far used domain of innovation-processes needs to be extended to be able to represent collaborative production planning processes that are built on the basis of distributed production planning processes. What actually are more detailed representations and therefore domain models for one of the phases of so-called innovation value chains. Therefore, activities in the generic collaborative production planning process need to be expressed in terms of distributed planning processes. That processes are annotated with resources, such as experts and formal representations of their tacit knowledge as well as documents capturing and bearing externalized knowledge. Future collaborative production planning processes will in this way be enabled to benefit from reusing and instancing these annotated generic planning processes as well as from underlying semantically annotated representations of planning activities and planning knowledge resources. The semantic schema of the WPIM application and the corresponding tool-suite is based on the ***Resource Description Framework (RDF)*** [11] and enables semantic-based searching by using the ***SPARQL Protocol And RDF Query Language (SPARQL)***. These enabling technologies provide a well-defined formal semantic description of knowledge. Using these explicit and machine readable representations of knowledge in distributed cross-organizational environments as known from the requirements of collaborations in the SME domain can improve collaboration between heterogeneous partners and add value to an advanced and even more integrated CAPP.

The WPIM application and the corresponding tool suite is using four layers for knowledge representation. It offers the opportunity to get on a top layer a brief overview of the innovation i.e., in the case of the CAPP planning process and if needed to navigate to deeper more detailed process descriptions, accompanying knowledge resources, documents as well as annotated attributes and features. The underlying ontology in the WPIM application and corresponding tool suite offers a machine-readable structure for concepts that can also be read and understood by human experts. Ontologies offer the opportunity to order concepts hierarchically as in e.g., a taxonomy but furthermore add non-hierarchical relationships between such concepts. For example, coming from a functional point of view for some applications the two concepts mechanical cutting and laser cutting can be understood as replaceable concepts. The ***Web Ontology Language (OWL)*** [12][13] allows to model concepts in classes and e.g., this replaceable relationship between these two classes of cutting technology. A production planner using semantic search/reasoning for cutting methods will find both options of cutting and also will get the hint that these two concepts can potentially substitute each other. In this way, representing such knowledge in a machine-readable semantic way can pave the way towards applying AI methods as can e.g., be build by means of automated semantic reasoning over semantic knowledge representations. Additionally, with the concepts of ***Master Processes*** (German: Masterprozess, ***MP***, see Figure 1), ***Process Instances*** (German: Prozessinstanz, ***PI***, see Figure 1) as well as **Activities** and **Tasks** the separation of modeling and capturing generic and instance specific (in the domain of CAPP, this means e.g., knowledge related to a certain machine vendor) knowledge is supported. In this way, the process artifact representation toolbox of WPIM allows reusing process steps and their associated knowledge in a seamless way.

III. WPIM IN THE DOMAIN OF PROCESS PLANNING

WPIM was originally developed to support innovation processes by providing existing innovation process knowledge in an explicitly represented form to innovation process experts as well to computer agents i.e., computing machines and their software programs. In the field of innovation processes, the usage and potential of semantically represented processes as enabled by WPIM has already been elaborated. Furthermore, WPIM has already been applied to represent PLM data in the field of technical products. In both domains next to executing processes also planning processes has been modeled and used for representation. Semantics as offered by WPIM have the advantage of being easily exchangeable and machine readable. This helps e.g., to plan cross-organizational and distributed innovation processes.

The following

Figure 1 describes the interaction of a MP with its PIs. If such processes need to be represented in WPIM, in a first step the user selects classes in the WPIM ontology repository to register an instance of a process resource. This means, the user e.g., selects the process classification systems to be used as the global set of ontologies into which the knowledge resource structure and contents are to be mapped. In a second step, the user selects attributes for each selected resource class for populating virtual objects in these classes with content resources. This implies, the user has also e.g., to map the attributes of the resources to specific ontologies. Thus, indicating that an attribute's contents (their range) is mapped to an ontology, such as mapping a resource attribute onto an expert ontology. Finally, the user selects the populating methods or populates the resource instances and their specific content manually.

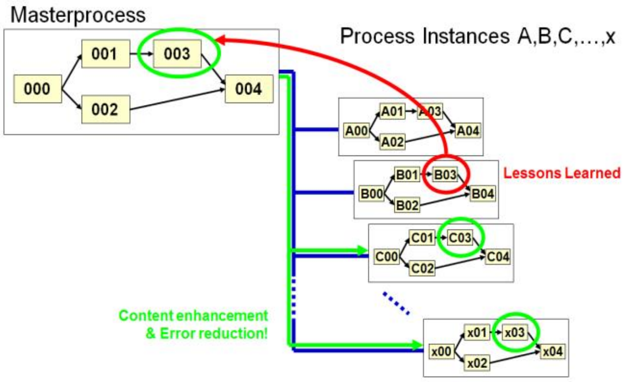


Figure 1. Master Process and Process Instances [2]

This means, the user maps the attributes of contents to classes in the ontology manually or semi-automatically using word-matching or other provided techniques e.g., map "hole" from a product property ontology concept to the "drilled hole" concept in the machining feature ontology.

However, before such mappings can be established the sources' local data schemas must first be registered. For example, in our implementation we used the two activity-based schemas displayed in Figure 1 for representing the MP and PI resources.

The next two sections describe in detail what an activity-based MP is and how activity-based executions of this MP (i.e., PIs) are defined.

A. *Master Processes*

A MP is a generic high-level description of a process. In WPIM, from a data set point of view, a MP describes a data structure and attributes of a higher level template for a process. The representation approach goes beyond the sole representation of the process structural schema but describes process structures and their attributes by using semantic representations. As WPIM offers such semantic descriptions of MPs, the semantic MP schema exists as a generic and formal description of a process, independent of generated data instances during a certain execution of the process. As an example, a MP defines next to a well-defined structure of contained activities. Resources, which will be involved during execution the process. For instance, this can be experts, documents or, in the case of a CAPP adaptation, could be production machines and their production activities.

B. Process, Activity and Task Instances

When executing a process, data is gathered. WPIM describes this, from the data set point of view, as a PI. The Activity structure that exists in WPIM and is displayed in

Figure 2 is used to store all outgoing and incoming data as well as Activity states. Beyond that, WPIM also allows to describe and represent PIs including their Activities in a semantic, machine-readable format. Furthermore, WPIM PIs are ordered in a chronological way. That means, if a first instance is e.g., executed, the Lessons Learned during that execution can be stored within the higher level MP and this gathered information can be provided for the following process execution within the next PI (see Figure 1).

An activity needs well defined inputs to generate a required output. Activities within WPIM contain one to many tasks. An instance of an Activity defines a cluster of tasks e.g., an Activity can bundle tasks that are assigned to a single resource. Such an assignment can contain planning tasks that need to be executed by an expert (e.g., a planner) or tasks can also be assigned to a resource like a machine in order to represent the execution of a machine operation.

In a WPIM context, a Task structure is an action that can-not be further split into sub-actions. WPIM offers a semantic data representation to archive status and values when performing a Task. Such a Task can for example, represent an operation that can be executed by a machine and create a specified result. By having such a semantic representation containing incoming and outgoing status, progress attributes and result specification, WPIM allows to delegate a Task instance to various executing entities. An example, in the context of planning tasks, it's to finalize a plan by signing the plan and setting it into action. A Signature to release a plan is a very unique task and it is obvious, that such a signing task cannot be split – either the plan is released via signature or it is not signed and therefore not released.

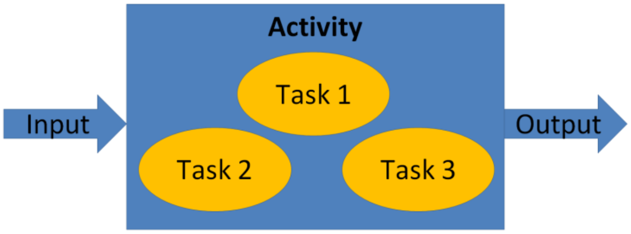


Figure 2. Visualization of an Activity as a set of Tasks

As displayed in

Figure 2, an Activity consists of at least one up to many Tasks. These Tasks represent the transformation of an input of the Activity into an output.

IV. Semantic Integration and Information Mediation within Knowledge-based Information System Architecture

Mediators are a standard approach in the construction of information system architectures. They have originally been introduced by Wiederhold in [14] as early as in 1991 when the web was still in its infancies and the semantic web did not even exist. However, since then, the use and application of these architectures in building web-based information systems supporting, data, information, and knowledge integration has grown into a de-facto standard. It is widely used in all types of scientific and industrial infrastructures supporting data, information, content and knowledge sharing, management and access for re-use. In the following, we will introduce the different levels of interoperability that can be addressed by mediator architectures in terms of integration. Furthermore, we will introduce markup languages as a means of defining global schemata and semantics for the purpose of semantic information integration and exchange. Finally, we will introduce mediator architectures of different types at increasing levels of detail supporting increasing levels of integration.

A. Levels of Interoperability and Integration

As outlined, e.g., in [15] data, information and knowledge integration can be understood at varying levels of interoperability and heterogeneity. In the following, we will describe this a bit more in detail based on a slightly adapted excerpt from [15]. When trying to share distributed and heterogeneous data, a number of technical challenges must be overcome. Consider, for example, two systems having data sets that should be made interoperable. One can employ standards and technologies to overcome the various kinds of heterogeneities and to facilitate interoperability at different levels. At the systems level, one may find different operating systems (Linux, MS Windows, MacOS, etc.), different data transport protocols (FTP or HTTP, which are built on top of a stack of internet protocols called TCP/IP etc.) or higher-level protocols for discovery and interoperation of web services. The Differences in system platforms and operating systems are usually overcome by standardizing protocols for data transport and remote service execution. For the latter, for example, one can employ web service descriptions (WSDL, 2001), which specify the input and output parameters of a web service. System level interoperability can also be achieved at the grid or cloud service level. Grid and cloud services extend the basic web-service infrastructure and include additional features such as user authentication for secure data access. Apart from the generic issues of data access, transport and remote execution, there are also a number of application specific system level issues e.g., the choice and architecture of the mapping technology for the integration and mediation of information and knowledge resources (server-side, client-side, mixed). At the syntactic level, one has to consider heterogeneities such as different data file formats, depending on the type of content or knowledge resource and corresponding representation format of the information and knowledge representation. The Extensible Markup Language (XML) [16] provides a simple and very flexible syntax for structuring many kinds of data, metadata, content and knowledge resources to enable their exchange. Defining such a new structure in XML syntax can be done in different ways. For example, one can provide an XML ***Document Type Definition*** (DTD) or an ***XML Schema Definition (XSD, XML Schema)*** [16][17] to specify the allowed nesting structure and (in XML Schema) the data types of XML elements.

In this way, XML not only yields a data, information, content and knowledge resource exchange syntax but also prescribes a schema for the exchanged resource. However, additional explicit representations of semantics such as domain specific integrity constraints have to be encoded by other means. The ***Resource Description Framework (RDF)*** [11] can be seen as an XML dialect for encoding labeled, directed graphs and in particular ontologies as an example of a standardized semantic vocabulary. For querying databases and query languages, such as the ***Standardized Query Language (SQL)*** [18] for relational databases) or ***XQuery*** (for XML databases) [19] are used, each of which come with their own syntax for query expressions. Differences at the syntactic level i.e., heterogeneity of the underlying data models of sources are usually resolved either by adhering to a standard or by using format converters that can translate from one format to another. At the schema level, heterogeneities can exist because the same (or at least similar) data can be represented using vastly different schema structures (even when the same file format or syntax is used). For example, two datasets may be organized in different ways across two relational databases i.e., the table and column structure may be very different although the content (at the conceptual level) of the databases may be very similar. Similarly, different DTDs or XML Schemas can be used to describe the same data for XML databases. To overcome schema level heterogeneities, we can again apply two approaches, schema standardization or schema transformation. For the latter, i.e., schema transformation, database query languages in general and XQuery in particular provide powerful means to express complex queries and transformations. Thus, (XML) query languages play an important role in database mediators. Finally, at the semantic level, we consider issues such as differences in terminology, different classification schemes and differences in the definition and constraints for the various concepts that are relevant to the data sets being integrated. Therefore, the main approach for reconciling semantic heterogeneities is the use of agreed-upon ontologies, which in their simplest form provide a controlled vocabulary with more or less formal descriptions of the pertinent concepts. In more sophisticated forms, ontologies include formalizations (often through logic formulas) of properties of concepts and "inter-dependencies" of concepts. A prominent emerging standard for ontologies is OWL, which comes in three increasingly expressive variants: ***OWL Lite, OWL DL and OWL Full*** [20]. OWL is also an interesting example of how several interoperability levels and standards may be intertwined: for example, OWL DL builds upon the RDF model and syntax which in turn is usually denoted in XML syntax.

B. Mediator Architectures

Database mediator systems can be used to provide uniform access to distributed heterogeneous data sets, and thereby overcome a number of the interoperability challenges mentioned above. Figure 3 depicts a typical mediator architecture in which a number of local data sources are "wrapped" as XML sources and subsequently combined into an integrated global view. Thus, a client application or the end user is provided with the illusion of querying a single, integrated (or global) database with one integrated schema.

Mediators are software components that serve to simplify, reduce, combine and explain data. They are mainly used for providing a common access level onto different distributed data sources. The source wrappers not only provide a uniform syntax, but also reconcile system aspects e.g., by means of a unified data access and query protocol [15].

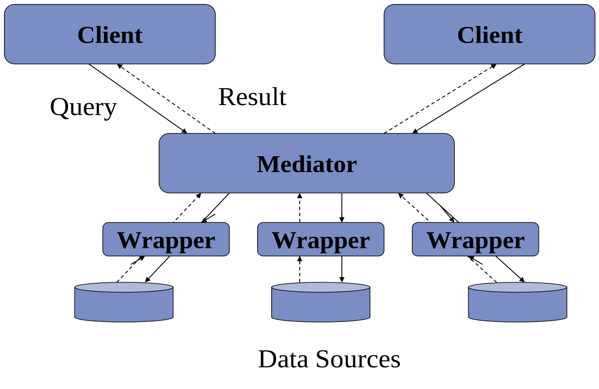


Figure 3. Mediator architecture integrating data sources

In a conventional relational or XML-based mediator system, interoperability is facilitated at the structural level. Differences in schema can be reconciled by corresponding schema transformation as part of the view definitions for the global view. However, terminological differences or other semantic differences are not adequately handled at the purely structural e.g., XML level. To this end, source schema and contents can be registered to an ontology, which encodes additional "knowledge" about the registered concepts. In the next section, we will explain more in detail how by means of "ontology-enabling" the system in this way can evaluate high-level queries over concepts that are not directly in the source databases and yet indirectly linked via an ontology. The task of the mediator is, to transform queries to the global schema into queries to the sources, to collect the results and to integrate and link them. The global scheme is based on a suitable data model, for which for example, XML or RDF can be used as representation. Wrappers are software components that represent the contents of a data source for the unification in another data model or schema. For example, XML wrappers are used to enable access to relational databases. The coupling between source and mediator via wrappers allows the mediator uniform access to the sources, by creating a mapping between the data model of the mediator and the data model of the local source. Also, incoming requests of the mediator can be translated into requests into the local source system.

C. Ontologies in Information Integration and Mediation

In information integration systems, based on a mediator architecture as displayed in Figure 4, ontologies can be used to provide information at the level of conceptual models and terminologies. Thereby, facilitating conceptual-level queries against sources and resolving some of the semantic-level heterogeneities between them. In our original WPIM system, the process classification ontology and the innovation ontology are used as a global view for registering process resources and processing queries. When a resource is registered to an ontology, a mapping from the data set to the selected ontology is generated. However, before such mapping can occur, the sources' local data schemas have to be registered first. After these steps, wrappers are created for the registered resources. Each wrapper uses the mappings between the data source and ontology to translate queries from the global ontology to the local schema and also to translate content from the local schema to the global ontology.

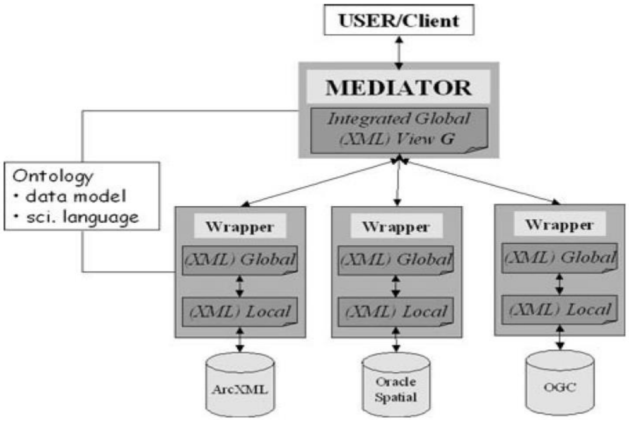


Figure 4. Extended Mediator Architecture [15]

As explained above, the system can automatically use the subclass relation to expand concept queries when required. Note that although all system-registered ontologies can be considered as conceptual-level query mechanisms, the system can suggest suitable ontologies based on: first, the user's choice of resources and second, the sources' schema information. Database mediator systems can be used to provide uniform access to distributed heterogeneous data sets [15].

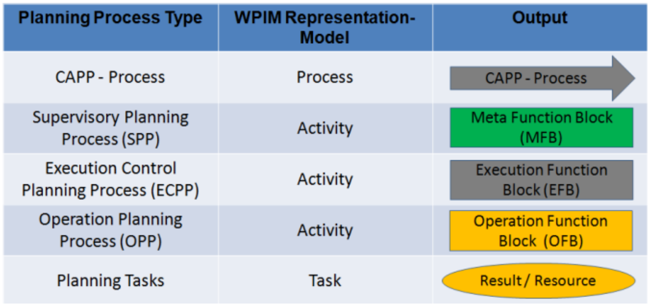
V. Overall CAPP Knowledge Integration and Mediation Concept

Beyond integrating data from distributed data sources, our proposed CAPP knowledge integration and mediation approach also describes how to combine heterogeneous data, information, content and knowledge sources by a mediation approach. For that approach, DPP process knowledge supporting CAPP activities need to be integrated by means of utilizing the WPIM semantic as well as need to be supported by mediation function, which allows integrated access through a global schema to the distributed CAPP knowledge resources in a DPP process. Over the course of the research on WPIM its field of application has been extended beyond just innovation management [21][22] and potentially can be applied as well in DPP and CAPP. This can be considered one of the phases of an innovation value chain. Furthermore, the web-based approach of the WPIM-Application-Tool suite supports working collaboratively in dispersed teams. In the domain of CAPP that means planning activities and consecutive manufacturing processes, which are handled by a network of many SMEs could benefit from such a common platform and therefore such use cases need further consideration.

A. *Collaborative Planning Processes*

CAPP processes aim to combine and integrate distributed information and knowledge resources e.g., about machine and tool descriptions, machine features and process constraints in order to create an executable plan for a certain task. Such CAPP activities can happen within the boundary of one organization or even across organizational boundaries. The CAPP-4-SMEs project [1] explicitly has defined the goal to research in the field of CAPP e.g., for the use case ***where Original Equipment Manufacturers (OEMs)*** work with global partners and suppliers, which are mainly SMEs, more collaboratively to achieve entire manufacturing value chain optimization [5]. This paragraph describes the concept of Collaborative Process Planning in CAPP-4-SMEs and the challenge of turning **the supervisory plan** into an **operational plan** in an optimized manner. The planning process approach to be used in CAPP-4-SMEs is a form of DPP. Most process plans generated using existing CAPP systems are tied to specific resources (machines, fixtures, cutters, etc.) and therefore are inflexible and not responsive to unexpected changes. The plan must be severely revised every time when a resource becomes unavailable. Which might mean that similar planning tasks have to be accomplished repetitively [23, p. 5]. The goal of DPP is to improve flexibility and adaptability and ultimately allow real-time manufacturing intelligence. Therefore, a process plan consists of two parts: While *generic data* (machining method, machining sequence and machining strategy) is used to describe one or many alternative plans (which then is called Non-Linear Process Planning [3, p. 54]) *machine-specific data* (tool, data, cutting conditions and tool paths) serves to choose from the actual resources available to produce the parts. This leads to a two-layer hierarchy, where the two different tasks can be accomplished at two different levels: shop-level SP and controller-level OP [23, p. 5ff].

To represent the derived planning information, the concepts of ***Machining Features (MFs)*** and FBs are used as enabling technologies. MFs typically represent shapes, which can be achieved by the available machining resources. As already described above, FBs are a concept provides control based on data flow and finite-state machine concept [23, p. 8ff]. The Decision Making for SP is non-trivial as there is not one single correct plan how to produce a part, as machining features applied in different sequences can be used to achieve the same result making non-linear process planning necessary. This task is covered in the steps machining sequence processing within the supervisory planning [23, p. 11ff]. In the following, and as a first step of semantic knowledge representation for the CAPP domain, this paper does focus on the semantic representation of SP and OP processes. These processes include cutter selection, operation sequencing, cutting parameter assignment and tool path generation. They vary on the basis of chosen machining strategy and machining dynamics that affect tool life and surface finish quality. Improper decisions at this level may result in tool breakage, chatter vibration and even scrap. The knowledge about choosing the right resources is either covered in vendor-specific handbooks or was gained through long-lasting experience of engineering experts who working for a specific company. That knowledge is either not extractable or therefore not representable in a standardized form (at least when looking at its informal encoding in handbooks) or even must be considered as implicit or tacit knowledge, when looking at the expert's experience. While, the ultimate goal of DPP is to do operation planning in an automated fashion adapting to available scheduling and availability monitoring information.



Planning Process Type	WPIM Representation-Model	Output
CAPP - Process	Process	CAPP - Process
Supervisory Planning Process (SPP)	Activity	Meta Function Block (MFB)
Execution Control Planning Process (ECPP)	Activity	Execution Function Block (EFB)
Operation Planning Process (OPP)	Activity	Operation Function Block (OFB)
Planning Tasks	Task	Result / Resource

Figure 5. CAPP Ontology based on WPIM Models and DPP Process Types and Resources/Results

The current reality is, that in many cases this planning step is still time and labor intensive and the required planning process themselves are not yet computer supported in terms of representing them in a machine readable semantic way. In the following, we will explain the representation, integration as well as mediation that can be achieved for representing CAPP activities based on DPP knowledge in an integrated way.

That is accessible on a global level although the DPP knowledge resources are coming from distributed sources of the collaborating agents/processes. Figure 5 outlines our integration approach that will further be elaborated in the following.

When combining, function blocks with WPIM we see strong advantages in both approaches. FBs are very planning oriented and focused on production domain. WPIM offers well described data structures for Processes, Activities and Tasks. In the following, we will now apply such WPIM process and resource representation structures, which are semantic-based and therefore give the possibility to represent data in an exchangeable, human-understandable and machine-readable format. For example, the created representation structure allows navigation from Process level to Activity and Task level and vice versa. In this way, the semantic representation structure will add value to distributing and at the same time sharing knowledge about production planning processes e.g., when exchanging single activities between processes and during allocation of tasks i.e., resources/ results to a machine level. In the understanding of WPIM, the DPP planning process and resource knowledge is represented by planning activities consuming and producing planning knowledge resources. These can e.g., be FBs over all levels of CAPP activities from ***SP Process (SPP)*** activities through ***ECP Process (ECPP)*** activities to ***OP Process (OPP)*** activities (see Figure 5). Therefore, a production of resulting planning results/resources from MFBs through EFBs to OFBs is possible. This process and resource knowledge can be brought into one integrated and well defined semantic schema with certain instances. In this way, the representation of the different types of planning activities producing and consuming FB resources by means of WPIM's semantic process representation schemas allows to represent a top down planning process representation schema. As well as a top-down mediation of different types of knowledge-resource and planning-result representations from higher levels of planning abstraction to lower levels of operational planning representation. In this way, WPIM provides an integrated and well-structured schema to be filled during execution with instances of semantic data on each level of planning abstraction and corresponding process and resource/result distribution.

As displayed in

Figure 6, a SPP can be represented by a WPIM Activity representation instance that transforms an input MFB on the basis of some additional planning resources produced by its tasks into an output MFB. Therefore, the EFB uses at least one EFB of an earlier iteration of a SPP activity.

This means, that the MFBs produced by the SPP activity as displayed in Figure 6 are not only consumed by future iterations of such an SPP activity but also get consumed by the underlying ECPP activity.

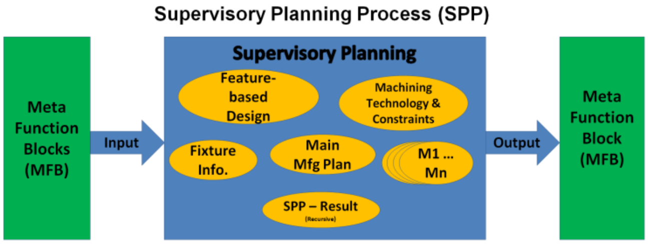

Supervisory Planning Process (SPP)
Meta Function Blocks (MFB)
Input
Supervisory Planning
Feature-based Design
Machining Technology & Constraints
Fixture Info.
Main Mfg Plan
M1 ... Mn
SPP – Result
(Recursive)
Output
Meta Function Block (MFB)

Figure 6. Supervisory Planning Process Activity

Also an ECPP can be represented by an instance of a WPIM activity as shown in Figure 7. This process transforms the incoming MFB provided by the SPP activity, the additional resource information (also MFBs) and the delivered OFB from the underlying OPP activity outgoing in an EFB. Therefore, an EFB uses at least one earlier iteration of a SPP activity and an OFB of the subsequent OPP activity. An ECPP activity (Figure 7) itself produces EFBs which get assigned to machines and consumed by them. In addition, the EFBs are used as inputs for the OCPP activities which are for producing and output of corresponding OFBs.

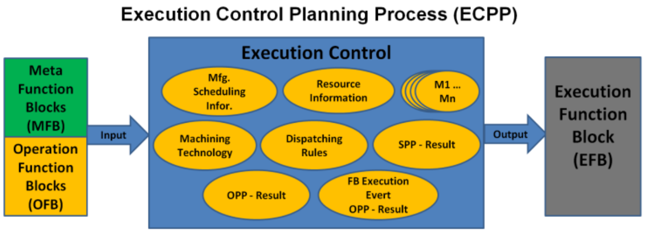

Execution Control Planning Process (ECPP)
Meta Function Blocks (MFB)
Operation Function Blocks (OFB)
Input
Execution Control
Mfg. Scheduling Infor.
Resource Information
M1 ... Mn
Machining Technology
Dispatching Rules
SPP - Result
OPP - Result
FB Execution Evert OPP - Result
Output
Execution Function Block (EFB)

Figure 7. Execution Control Planning Process Activity

Analogously to the first two, an OPP can also be represented by an instance of WPIM activity representation. The OPP activity (Figure 8) transforms the already explained EFB that created apriori from the ECPP activity as well as several other information, like status and events (all MFBs) in an outbound OFB. Furthermore, in the DPP methodology OFBs have a direct link to the real execution of the process. That means, that OFBs are executed by a directly assigned resource e.g., a machine that at the same time produces a certain result in this way that can be re-used as a resources in the remainder of the planning process.

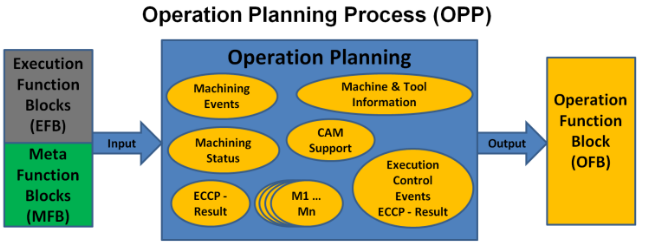

Operation Planning Process (OPP)
Execution Function Blocks (EFB)
Meta Function Blocks (MFB)
Input
Operation Planning
Machining Events
Machine & Tool Information
Machining Status
CAM Support
ECCP - Result
M1 ... Mn
Execution Control Events ECCP - Result
Output
Operation Function Block (OFB)

Figure 8. Operation Planning Process Activity

To achieve a representation of this, this kind of subprocess structure on the basis of WPIM, the Process Planning levels ECPP and OPP have to be represented as additional underlying WPIM activities of the same MP. Therefore, the resulting outputs of these processes (EFBs and OFBs) have been represented as planning results and therefore as knowledge resources that are handed over between these three planning activity levels of the same overall DPP MP. In summary, this means that the whole DPP methodology as applied in CAPP application domains can be represented by WPIM as a three level integrated WPIM activity. That representation belongs to one overall DPP MP where the WPIM Activities represent SPPs, ECPPs and OPPs as well as their results/resources which are tasks for the activities itself.

However, besides an integration on the level of the knowledge representation the WPIM system also needs to be extended to support access to distributed resources of such potentially distributed planning processes from a system distribution point of view. Therefore we conclude our approach in the following with a corresponding design of a three level mediator architecture that can handle the above described process and resource representations.

VI. EXTENDING WPIM TO INTEGRATE DPP KNOWLEDGE AND MEDIATE ITS ACCESS DURING CAPP

Figure 9 displays a first level mediator architecture that integrates MFBs and other relevant and potentially distributed resources for the SPP activity from the different levels of the overall CAPP process that is implemented by means of the DPP method. The resulting mediator is called the SPP Mediator.

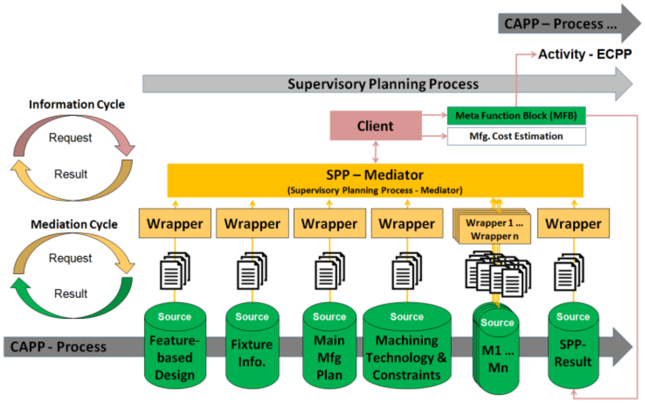


Figure 9. First level Mediator Architecture using for SPP

Therefore, a down-stream DPP mediation can be implemented by means of two analogously derived additional mediators on the second and the third DPP level.

On the second level of the mediator architecture follows then the deduced and so-called ECPP mediator which supports the above-mentioned ECPP activity. Figure 10 shows this second level of the mediator architecture. They assimilated at least an earlier iteration of the SPP-mediator as MFB and a OFB of the subsequent OPP mediator (level 3) and various other relevant and potentially distributed resources.

Coming from the machining-data point of view, the corresponding up-stream Mediation Process starts from machines with a defined need of steering information which can be harmonized by using wrappers and offering a mediated interface to clients.

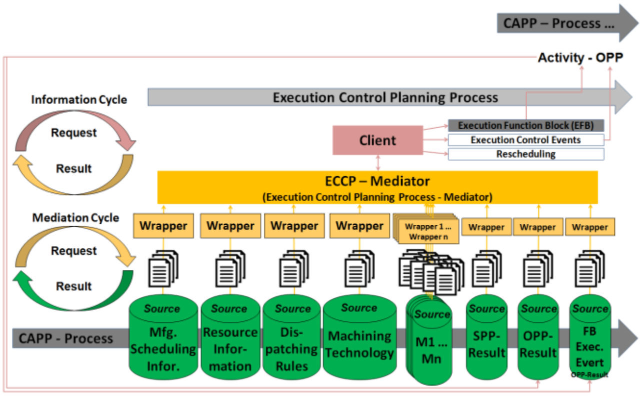


Figure 10. Second level Mediator Architecture using for ECPP

The third and final level of the mediator architecture of the CAPP process forms the again derived OPP mediator. Figure 11 represents this level graphically and displays how the so-called OPP mediator completes the mediation process. This integrates relevant and potentially distributed machine resources as MFBs and by the second level generated EFBs (ECPP-mediator) for the OPP activity. This three-tier architecture can support an Information Process by, providing data from distributed data repositories, combining various data formats, in a single semantic enabled format, as well as a mediation process requesting, accessing and collecting/gathering/combining data from different distributed resources.

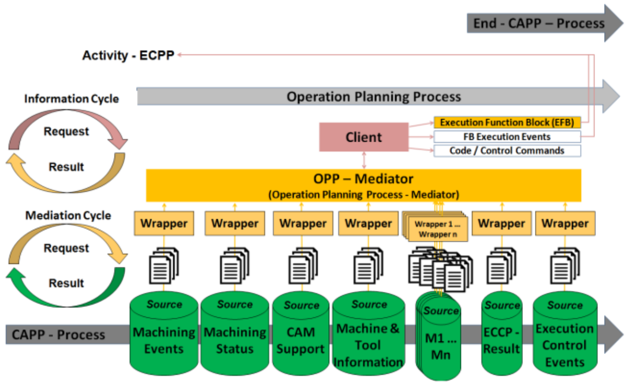


Figure 11. Third level Mediator Architecture using for OPP

The appendix contains a detailed illustration of the entire CAPP process of the mediator architecture (Figure 12) to get a good overall understanding and to clarify the relationships and dependencies between the individual levels of mediation.

In summary, this means that DPP i.e., deriving the Operational Plan from the Supervisory Plan through the Execution Control Plan is therefore a three-level WPIM Process where the three levels can be modeled as interlinked WPIM activities.

VII. Conclusion and Future Work

This paper has presented the relevant state-of-the-art and derived a method to support semantic knowledge management of DPP knowledge in the CAPP application domain based on semantic process representations producing and consuming function blocks and other relevant planning resources for distributed production planning. In this way, the challenges of planning resource distribution, sharing and mediation that are inherent to CAPP are addressed in a first initial step of modeling this domain. Besides this, requirements towards representing this knowledge in a machine readable way on the one hand and on the other hand designing an implementation architecture that can deploy such a CAPP support into a cloud-based i.e., highly distributed or even fully virtualized system distribution are proposed.

In this way, our approach will allow e.g., SMEs to participate in a cloud based CAPP activity that is implemented on the basis of the DPP method. This is represented by the WPIM methodology in a machine readable way and where the distribution architecture within the cloud and beyond is achieved on basis of applying a three level mediator architecture. By extending the WPIM system with such a three level resource mediation architecture, users will be enabled to create process instances of the provided DPP master processes representing all three levels of the DPP planning process activities and all their resources and results from the highest level of product features down to the lowest level of machining features. By doing so, the individual SMEs can reflect, which resources they have available and can annotate the DPP knowledge representation they have received and in this way documenting their potential competitive advantage. With this approach, we see the potential to address several issues existing today.

Firstly, on a general level - to capture DPP knowledge needed for process planning current tools are still rather complex to maintain and therefore not every SME has the capacity to run and maintain such a system. By delivering this functionality through a cloud-based repository approach building on semantic–web enabled knowledge representations and integration as well as mediation support. The usage of such tools can be provided at an affordable usage fee.

Secondly, by the ability to provide knowledge not specific to a certain company or vendor of machines via e.g., a subscription model that is enabled through such an approach. SMEs, which do not have the manpower to build up that knowledge within their own research and engineering organization can source out this generic CAPP knowledge and start directly on enhancing their specific DPP knowledge increasing their competitive advantage in their respective production support niche. On a more specific level this approach fosters two aspects: From a knowledge management point of view the existence of explicit knowledge being available through handbooks etc. is made visible in a consistent and machine readable manner. The other fact is that tacit knowledge exists within the minds of long-standing employees is externalized by annotating these persons to specific process steps as expert. Referencing the SECI model [21, p. 20] this can be used for knowledge conversion through socialization (based on the annotation in the WPIM process colleagues start asking questions to the experts about that matter and the tacit knowledge gets spread). From a collaboration aspect, this approach can support teams within a company and beyond the borders of an organization to collaboratively improve planning results. They can trigger knowledge conversion through socialization across the boundary of different sites of a company, which unlikely would happen if the fact that tacit knowledge exists (even though not the knowledge itself) would not be externalized. While supporting such a scenario within one company can be beneficial it would also be beneficial when several companies do work together in a manufacturing network.

As a further development of this work, our next step is the practical implementation. For this purpose, we need more typical examples for the three step mediator architecture and we want to reimplement and extend the WPIM tool suite. Thus, the theoretical preparatory work will be also practically applied and implemented.

Acknowledgment and Disclaimer

This publication has been produced in the context of the CAPP-4-SMEs project. The CAPP-4-SMEs project has received funding from the European Union's Seventh Framework Programm for re-search, technological development and demonstration under grant agreement no 314024.
However, this paper reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains;

References

[1] F. Miltner, T. Vogel and M. Hemmje, "Towards Knowledge Based Process Planning Support for CAPP-4-SMEs: Problem Description, Relevant State of the Art and Proposed Approach", vol. 1, International Manufacturing Science and Engineering Conference (MSEC), 2014.

[2] T. Vogel, "Wissensbasiertes und Prozessorientiertes Innovationsmanagement WPIM - Innovationsszenarien, Anforderungen, Modell und Methode,Implementierung und Evaluierung anhand der Innovationsfähigkeit fertigender Unternehmen", Dissertation, Hagen, 2012.

[3] L. Wang, G. Adamson and M. H. a. P. Moore, "A Review of Function Blocks for Process Planning and Control of Manufacturing Equipment", Journal of Manufacturing Systems, Vol.31, No.3, pp.269-279, 2012.

[4] L. Wang, W. Jin and H. Y. Feng, "Embedding machining features in function blocks for distributed process planning," *International Journal of Computer Integrated Manufacturing,* pp. 443-452, 2006.

[5] L. Wang, H. Y. Feng and N. Cai, "Architecture design for distributed process planning," *Journal of Manufacturing Systems*, pp. 99-115, 2003.

[6] M. Helgoson, L. Wang, R. Karlsson, M. Givehchi and M. Tedeborg, "Concept for Function Block enabled Process Planning towards multi-site Cloud Collaboration", International Manufacturing Science and Engineering Conference (MSEC), Vol. 1, 2014.

[7] International Electrotechnical Commission, Switzerland: Function blocks – Part 1: Architecture, IEC 61499-1, 2005.

[8] R. Lewis, "Modelling control systems using IEC 61499 – applying function blocks to distributed systems", ISBN: 0852976 796: The Institution of Electrical Engineers, 2001.

[9] International Electrotechnical Commission, Switzerland: Programmable controllers – Part 3: Programming languages, IEC 61131-3, 2003.

[10] T. Vogel and M. Hemmje, "Auf dem Weg zu einem Wissens-basierten und Prozess-orientierten Innovationsmanagement (WPIM) – Innovations-szenarien, Anforderungen und Modellbildung," in *KnowTech 2006*, Poing, CMP-WEKA-Verlag, 2006.

[11] R. Cyganiak, D. Wood, M. Lanthaler, G. Klyne, J. Carroll and B. McBride, "RDF 1.1 Concepts and Abstract Syntax," W3C Recommendation 25 February 2014, World Wide Web Consortium (W3C), http://www.w3.org/TR/rdf11-concepts/,, Feb 2014, last accessed Nov 13, 2014.

[12] W3C OWL Working Group, "OWL 2 Web Ontology Language Document Overview (Second Edition)," W3C Recommendation 11 December 2012, World Wide Web Consortium (W3C), http://www.w3.org/TR/owl2-overview/, December 2012, last accessed Nov 13, 2014.

[13] W3C, "OWL Web Ontology Language Overview,," World Wide Web Consortium,, 10 February 2004. [Online]. Available: http://www.w3.org/TR/owl-features/. [Accessed 14 November 2013].

[14] G. Wiederhold, "Mediators in the Architecture of Future Information Systems", The IEEE Computer Magazine, 1992.

[15] B. Ludäscher, K. Lin, B. Brodaric and C. Baru, "GEON: Toward a Cyberinfrastructure for the Geosciences—A Prototype for Geologic Map Integration via Domain Ontologies", Digital Mapping Techniques '03 — Workshop Proceedings, U.S. Geological Survey Open-File Report 03–471, 2003.

[16] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler and F. Yergeau, "Extensible Markup Language (XML) 1.0 (Fifth Edition), W3C Recommendation 26 November 2008, World Wide Web Consortium (W3C)," http://www.w3.org/TR/REC-xml/, last accessed Nov 2014, November 2008.

[17] S. Gao, C. M. Sperberg-McQueen, H. S. Thompson, N. Mendelsohn, D. Beech and M. Maloney, ""W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures", W3C Recommendation 5 April 2012, World Wide Web Consortium (W3C), http://www.w3.org/TR/xmlschema11-1/," last accessed Nov 13, 2014, 5 April 2012.

[18] J. Melton, "ISO/IEC FDIS 9075-1 Information technology - Database languages - SQL - Part 1: Framework (SQL/Framework), ISO Draft International Standard, ISO/IEC JTC 1/SC 32 Data Management and Interchange," http://www.jtc1sc32.org/doc/N2151-2200/32N2153T-text_for_ballot-FDIS_9075-1.pdf, last accessed Nov 13, 2014, August 2011.

[19] J. Robie, D. Chamberlin, M. Dyck and J. Snelson, "XQuery 3.0: An XML Query Language," W3C Recommendation 08 April 2014, World Wide Web Consortium (W3C), http://www.w3.org/TR/xquery-30/, April 2014.

[20] B. Motik, B. Cuenca Grau, I. Horrocks, Z. Wu, A. Fokoue and C. Lutz, "OWL 2 Web Ontology Language Profiles (Second Edition)," W3C Recommendation 11 December 2012, World Wide Web Consortium (W3C), http://www.w3.org/TR/owl2-profiles/, last accessed Nov 13, 2014, December 2012.

[21] I. Nonaka and D. J. Teece, Managing Industrial Knowledge: Creation, Transfer and Utilization, London: SAGE Publications, 2001, pp. 13-28.

[22] F. Miltner, "Wissensbasiertes Prozessmanagement - Rollen, Kollaborationen und Schnittstellen - am Beispiel der Integration von SharePoint und WPIM," Hagen, 2013.

[23] L. Wang and W. Shen, Process planning and scheduling for distributed manufacturing, London: Springer, 2007.

[24] Environmental Systems Research Institute Inc., "Technical Description, An ESRI White Paper," http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf, Redlands, CA, July 1998.

[25] ISO International Standard 10303-1:1994 Industrial automation systems and integration -- Product data representation and exchange -- Part 1: Overview and fundamental principles, International Organization for Standardization, Geneva, Switzerland (1994).

[26] ISO International Standard 10303-11:1994, Industrial automation systems and integration — Product data representation andexchange — Part 11: Description methods: The EXPRESS language

reference manual, International Organization for Standardization, Geneva, Switzerland (1994).

[27] ISO International Standard 14649-1:2003. Industrial automation systems and integration -- Physical device control -- Data model for computerized numerical controllers -- Part 1: Overview and fundamental principles. Geneva: International Organization for Standardization. Retrieved 2008-10-27.

[28] Vogel, Tobias. www.inKNOWvation.de. WPIM - Wissensbasiertes und Prozessorientiertes Innovationsmanagement. [Online] 2012. [Quote from: 1st May 2015.] http://www.inKNOWvation.de.

APPENDIX

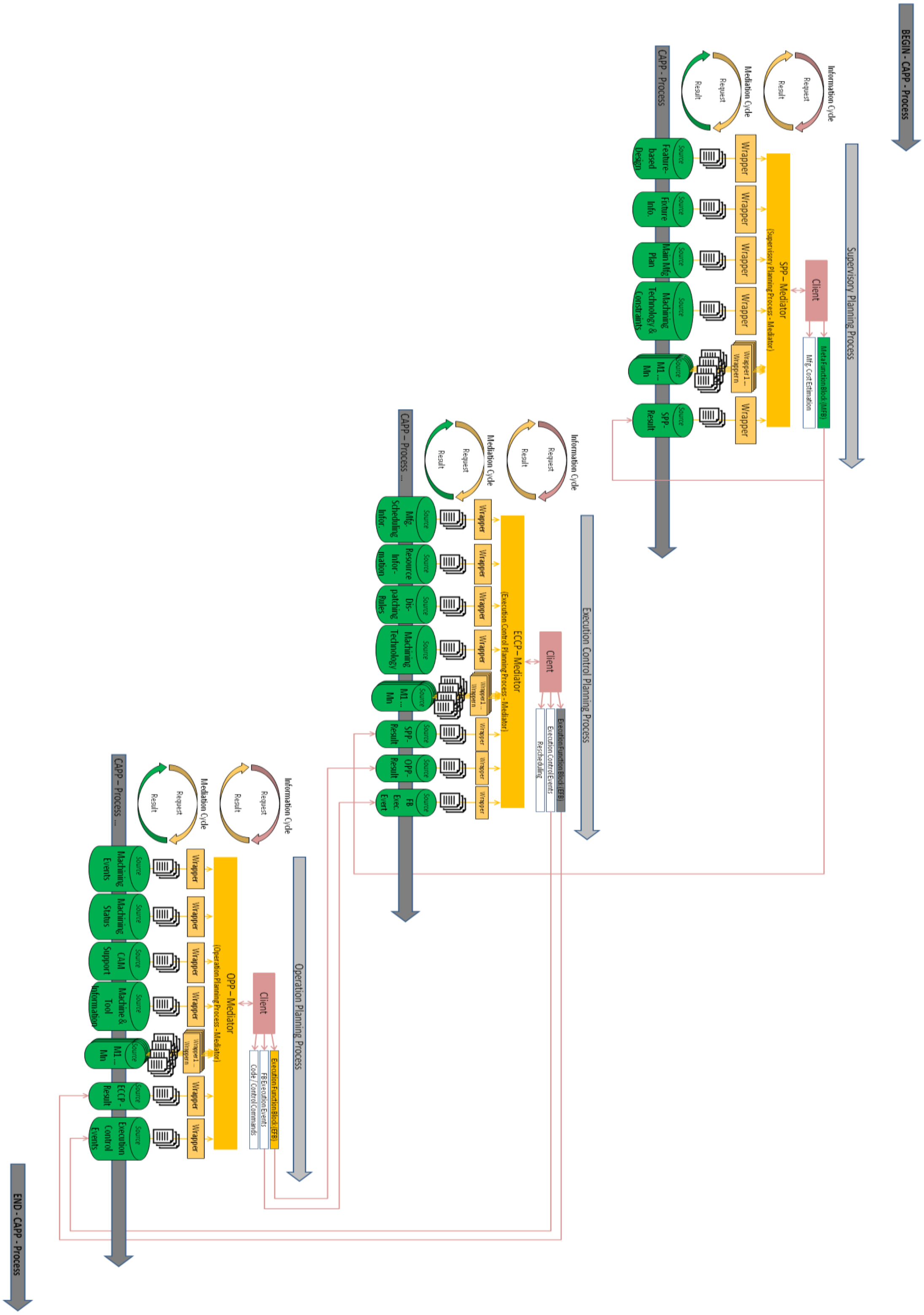


Figure 12. Entire CAPP process mediator architecture

On The Idle Time Model In Computer Networks

Saulius Minkevičius
VU Institute of Mathematics and Informatics
Akademijos 4, 08663 Vilnius, Lithuania
and Vilnius University, Naugarduko 24, 03225
Vilnius, Lithuania
e-mail: minkevicius.saulius@gmail.com

Edvinas Greičius
Vilnius University, Naugarduko 24, 03225 Vilnius,
Lithuania
e-mail: edvinas.greicius@gmail.com

Abstract – **An open queueing network model in light traffic and heavy traffic has been developed. The probability limit theorem for the idle time process of customers has been presented in light traffic in open queueing networks. Also, we present an application of the theorem - an idle time model from the computer network practice.**

Keywords – mathematical models of technical systems; performance evaluation; reliability theory; queueing theory; open queueing network; light traffic; heavy traffic.

I Introduction and statement of the problem

The modern queueing theory is one of the most powerful tools for a quantitative and qualitative analysis of communication systems, computer networks, transportation systems, and many other technical systems. In this paper, we analyze queueing systems, arising in the network theory and communications theory (called an open queueing network). In this paper, probability limit theorems are considered by investigating the values of a virtual time of a customer and an idle time of a customer in an open queueing network.

One of the main directions of research in the theory of queues corresponds with the asymptotic analysis of formulas or equations, describing the distribution of this and another probabilistic characteristic of a queue. For such a development of the analysis, there must be formulas or equations and besides that an unlimited convergence of the queue to the critical own point. This way, the first results about limited behaviour of single channel queues in heavy traffic are achieved (see [6, 7]). In the single phase case, where intervals of times between the arrival of customers are independent identically distributed random variables and there is one single device, working independently of the output in heavy traffic, is fully investigated in well known papers [2, 3]. Probability limit theorems for a virtual waiting time of a customer in a single queue are proved under various conditions of heavy traffic (see [8]). Probability limit theorems for a virtual waiting time of a customer in heavy traffic and an idle time of a customer in light traffic are closely connected. Also probability limit theorems for the waiting time of a customer and the queue length of customers in a multiphase queue are proved under various conditions of heavy traffic (see [10]).

So, in the sequel, we present a probability limit theorem in conditions of light traffic for another important probabilistic characteristic of an open queueing network (idle time of a customer). Note that there are only some works designed for the investigation of the idle time of a customer in a single-server queue (see surveys [16, 18] and paper [17]). Also, note that the research of the idle time of a customer in more general systems than the classical system $GI/G/1$ (multiserver queue, multiphase queue, open queueing network, etc.) has just started (see again [16, 17, 18]).

The idle function of computer networks shows in which part of time the computer network is not busy (idle). So in this paper, we present the probability limit theorem for the idle time of a customer in light traffic in the queueing network. The service discipline is “first come, first served” (FCFS). We consider the open queueing network with the FCFS service discipline at each station and general distributions of interarrival and service time. The queueing network studied has k single server stations, each of which has an associated infinite capacity waiting room. Each station has an arrival stream from outside the network, and the arrival streams are assumed to be mutually independent renewal processes. Customers are served in the order of arrival and after service they are randomly routed either to another station in the network, or out of the network entirely. Service times and routing decisions form mutually independent sequences of independent identically distributed random variables. The

ISBN: 978-1-61208-432-9

basic components of the queueing network are arrival processes, service processes, and routing processes.

We begin with a probability space (Ω, B, P) on which these processes are defined. In particular, there are mutually independent sequences of independent identically distributed random variables $\left\{z_n^{(j)}, n \geq 1\right\}$, $\left\{S_n^{(j)}, n \geq 1\right\}$ and $\left\{\Phi_n^{(j)}, n \geq 1\right\}$ for $j = 1, 2, ..., k$; defined on the probability space. The random variables $z_n^{(j)}$ and $S_n^{(j)}$ are strictly positive, and $\Phi_n^{(j)}$ have a support in $\{0, 1, 2, ..., k\}$.

We define $\mu_j = \left(E\left[S_n^{(j)}\right]\right)^{-1}$, $\sigma_j = D\left(S_n^{(j)}\right)$ and $\lambda_j = \left(E\left[z_n^{(j)}\right]\right)^{-1}$, $a_j = D\left(z_n^{(j)}\right)$, $j = 1, 2, ..., k$; with all of these terms assumed finite. Denote $p_{ij} = P\left(\Phi_n^{(i)} = j\right)$, $i, j = 1, 2, ..., k$. The $k \times k$ matrix $P = (p_{ij})$ is assumed to have a spectral radius strictly smaller than a unit. The matrix P is called a routing matrix. In the context of the queueing network, the random variables $z_n^{(j)}$ function as interarrival times (from outside the network) at the station j, while $S_n^{(j)}$ is the nth service time at the station j, and $\Phi_n^{(j)}$ is a routing indicator for the nth customer served at the station j. If $\Phi_n^{(i)} = j$ (which occurs with probability p_{ij}), then the nth customer, served at the station i, is routed to the station j. When $\Phi_n^{(i)} = 0$, the associated customer leaves the network.

At first, let us define $I_j(t)$ as the idle time of a customer at the jth station of the queueing network in time t (time t at which an open queueing network is not busy (idle) serving customers at the jth station of the queueing network), $\beta_j = 1 - \frac{\lambda_j + \sum_{i=1}^{k} \mu_i \cdot p_{ij}}{\mu_j}$, $\hat{\sigma}_j^2 = \sum_{i=1}^{k} p_{ij}^2 \cdot \mu_i \cdot \left(\sigma_j + \left(\frac{\mu_i}{\mu_j}\right)^2 \cdot \sigma_i\right) + \lambda_j \cdot \left(\sigma_j + \left(\frac{\lambda_j}{\mu_j}\right)^2 \cdot a_j\right)$, $j = 1, 2, \cdots, k$ and $t > 0$.

We suppose that the following conditions are fulfilled:

$$\lambda_j + \sum_{i=1}^{k} \mu_i \cdot p_{ij} < \mu_j, \ \ j = 1, 2, \ldots, k. \tag{1}$$

In addition, we assume throughout that

$$\max_{1 \leq j \leq k} \sup_{n \geq 1} E\left\{\left(z_n^{(j)}\right)^{2+\gamma}\right\} < \infty \ \text{ for some } \ \gamma > 0, \tag{2}$$

$$\max_{1 \leq j \leq k} \sup_{n \geq 1} E\left\{\left(S_n^{(j)}\right)^{2+\gamma}\right\} < \infty \ \text{ for some } \ \gamma > 0. \tag{3}$$

Conditions (2) and (3) imply the Lindeberg condition for the respective sequences.

One of the results of the paper is the following probability limit theorem for the idle time of a customer in an open queueing network (the proof can be found in [12]).

Theorem 1. *If conditions (1) - (3) are fulfilled, then*

$$\lim_{n \to \infty} P\left(\frac{I_j(nt) - \beta_j \cdot n \cdot t}{\hat{\sigma}_j \cdot \sqrt{n}} < x\right) = \int_{-\infty}^{x} \exp(-y^2 \cdot t/2) dy,$$

$0 \leq t \leq 1$ *and* $j = 1, 2, \ldots, k$.

II Idle time function of a computer network

Now we present a technical example from the computer network practice. Assume that queues of customers requests arrive at the computer v_j at the rate λ_j per hour during business hours, $j = 1, 2, \ldots, k$. These queues are served at the rate μ_j per hour by the computer v_j, $j = 1, 2, \ldots, k$. After service by the computer v_j, with probability p_j (usually $p_j \geq 0.9$), they leave the network and with probability p_{ji}, $i \neq j$, $1 \leq i \leq k$ (usually $0 < p_{ji} \leq 0.1$) arrive at the computer v_i, $i = 1, 2, \ldots, k$. Also, we assume the computer v_j to be idle, when the idle time of the waiting for service computer is less than k_j, $j = 1, 2, \ldots, k$.

In this section, we prove the following theorem on the idle time function of the computer network (probability of idle time in a computer network).A computer network is idle when it is not busy.

Theorem 2. *If* $t \geq \max_{1 \leq j \leq k} \frac{k_j}{\hat{\beta}_j}$ *and conditions (1) - (3) are fulfilled, all computers in the network are idle.*

Therefore, using Theorem 1, we get

$$\lim_{n \to \infty} P\left(\frac{I_j(n) - \beta_j \cdot n}{\hat{\sigma}_j \cdot \sqrt{n}} < x\right) = \int_{-\infty}^{x} \exp(-y^2/2) dy, \ j = 1, 2, \ldots, k. \tag{4}$$

Let us investigate a computer network which consists of the elements (computers) v_j, $j = 1, 2, \ldots, k$.

Denote

$$X_j = \begin{cases} 1, & \text{if the element } v_j \text{ is idle} \\ 0, & \text{if the element } v_j \text{ is not idle,} \end{cases}$$

$j = 1, 2, \ldots, k$.

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-432-9

Note that $\{X_j = 1\} = \{I_j(t) < k_j\}, \quad j = 1, 2, \ldots, k$.

Assume the structural function of the system of elements is connected with scheme 1 from k (see, for example, [2]) as follows:

$$\phi(X_1, X_2, \ldots, X_k) = \begin{cases} 1, & \sum_{i=1}^k X_i \geq 1 \\ 0, & \sum_{i=1}^k X_i < 1. \end{cases}$$

Suppose $y = \sum_{i=2}^k X_i$. Let us estimate the idle function of the system (computer network) using the formula of the full conditional probability

$$h(X_1, X_2, \ldots, X_k) = E\phi(X_1, X_2, \ldots, X_k) =$$

$$P(\phi(X_1, X_2, \ldots, X_k) = 1) = P(\sum_{i=1}^k X_i \geq 1) =$$

$$P(X_1 + y \geq 1) = P(X_1 + y \geq 1 | y = 1) \cdot P(y = 1) +$$

$$P(X_1 + y \geq 1 | y = 0) \cdot P(y = 0) = P(X_1 \geq 0) \cdot P(y = 1) +$$

$$P(X_1 \geq 1) \cdot P(y = 0) \leq P(y = 1) + P(X_1 \geq 1)$$

$$= P(y = 1) + P(X_1 = 1) \leq P(y \geq 1) + P(X_1 = 1)$$

$$= P(\sum_{i=2}^k X_i \geq 1) + P(X_1 = 1) \leq \cdots \leq$$

$$\leq \sum_{i=1}^k P(X_i = 1) = \sum_{i=1}^k P(I_i(t) \leq k_i).$$

Thus,

$$0 \leq h(X_1, X_2, \ldots, X_k) \leq \sum_{i=1}^k P(I_i(t) \leq k_i). \tag{5}$$

Applying Theorem 1 (with $t = 1$), we obtain that

$$0 \leq \lim_{t \to \infty} P(I_j(t) < k_j) = \lim_{n \to \infty} P(I_j(n) < k_j) =$$

$$\lim_{t \to \infty} P\left(\frac{I_j(n) - \beta_j \cdot n}{\hat{\sigma}_j \cdot \sqrt{n}} < \frac{k_j - \beta_j \cdot n}{\hat{\sigma}_j \cdot \sqrt{n}}\right) =$$

$$\int_{-\infty}^{-\infty} \exp(-y^2/2) dy = 0. \tag{6}$$

Then (see (6)),

$$\lim_{t \to \infty} P\left(I_j(t) < k_j\right) = 0, \; j = 1, 2, \ldots, k. \tag{7}$$

So, $h(X_1, X_2, \ldots, X_k) = 0$. (see (5) and (7)). The proof of the theorem is completed.

Further, we suppose that the following alternative conditions are fulfilled:

$$\lambda_j + \sum_{i=1}^k \mu_i \cdot p_{ij} > \mu_j, \; j = 1, 2, \ldots, k. \tag{8}$$

Let us define $V_j(t)$ as a virtual waiting time of a customer at the jth station of the queueing network in time $t, \; j = 1, 2, \ldots, k$. Note that this condition guarantees that, with probability one, there exists a virtual waiting time of a customer and this virtual waiting time of a customer when is constantly growing. One of the results of the paper is the following theorem on the probability limit for the virtual waiting time of a customer in an open queueing network.

Theorem 3. *If conditions (8) are fulfilled, then*

$$\lim_{n \to \infty} P\left(\frac{V_j(nt) - \beta_j \cdot n \cdot t}{\hat{\sigma}_j \cdot \sqrt{n}} < x\right) = \int_{-\infty}^{x} \exp(-y^2/2) dy,$$

$0 \leq t \leq 1$ *and* $j = 1, 2, \ldots, k$.

Proof. This theorem is proved under conditions $\lambda_j > \mu_j, \; j = 1, 2, \ldots, k$ (see, for example, [12]). Applying the methods of [15], it can be proved that this theorem is true under more general conditions(8). The proof of the theorem is complete.

Applying Theorem 3, we get the following result.

Theorem 4. *If* $t \geq \max_{1 \leq j \leq k} \frac{k_j}{\hat{\beta}_j}$ *and conditions (1), (2), and (8) are fulfilled, the computer network becomes unreliable (all computers fail). So, in that case, all the computers are busy.*

Proof. The proof is similar to that of Theorem 2. The proof of the theorem is complete.

Finally, we find the exact expression for $h(X_1, X_2, \ldots, X_k), \; t > 0$. Next, we prove the following theorem on this probability.

Theorem 5 $h(X_1, X_2, \ldots, X_k)$ *is equal to* $\exp(-\sum_{j=1}^k P(I_j(t) < k_j))$.

Proof. First denote $\lambda_j, \; j = 1, 2, \ldots, k$ as intensivities of structural elements, that form a complex stochastic system. Then the probability of stopping this system is equal to $e^{-\sum_{j=1}^k \lambda_j}$.

However,

$$\lambda_j = MX_j = P(X_j = 1) = P(I_j(t) < k_j), \; j = 1, 2, \ldots, k. \tag{9}$$

Applying (8), we obtain that $h(X_1, X_2, \ldots, X_k)$ is equal to

$$e^{-\sum_{j=1}^k \lambda_j} = e^{-\sum_{j=1}^k P(V_j(t) < k_j)}.$$

The proof is complete.

As a result, using Theorem 5, it is possible to estimate the idle time of a complex computer network.

III Concluding remarks and future research

1.1. Conditions (1)-(3) mean that the number of jobs arriving at the node of the network is greater than the number of service jobs at the same node of the network. It is clear from this note, that the length of jobs in the node of the network is constantly growing with probability one.
1.2. Conditions (1), (2), and (8) mean that the total number of jobs arriving at the node of the network is less than the number of service jobs at the same node of the network.It is clear from this note, that the length of jobs in the node of the network is constantly decreasing with probability one.
2. Now all the cases of traffic in computer networks are investigated – light traffic (see Theorem 1), average traffic (see Theorem 3) and heavy traffic (see Theorem 2). The investigation of all these cases is new.
3. If the conditions of the Theorem 1 and Theorem 4 are fulfilled (i. e., conditions (1) and (8) are satisfied), the network is either idle or busy. Conditions (1) and (8) are fundamental, - the behaviour of the whole network and its evolution is not clear, if conditions (1) and (8) are not satisfied. Therefore, this fact is the object of further research and discussion.
4. The theorems of this paper are proved for a class of open queueing networks in light traffic with the service principle FCFS, endless waiting time of customers at the each node of the queueing system, and the intervals between the arrival of customers at the open queueing networks are independent identically distributed random variables. However, similar theorems can be applied to a wider class of open queueing networks in the light traffic: when the arrival and service of customers in a queue are distributed in groups, also, when interarrival times of customers at the open queueing network are weakly dependent random variables, etc.

IV Acknowledgement

This research was partly supported by the National Complex Programme “Theoretical and engineering aspects of e-service technology creation and application in high-performing calculation platforms”.

References

[1] P. Billingsley, *Convergence of probability measures*, Wiley, New York (1968).

[2] A. Borovkov, *Stochastic processes in queueing theory*, Nauka, Moscow (1972) (in Russian).

[3] A. Borovkov, *Asymptotic methods in theory of queues*, Nauka, Moscow (1980) (in Russian).

[4] D. Iglehart, Multiple channel queues in heavy traffic. IV. Law of the iterated logarithm, Zeitschrift für Wahrscheinlicht-Keitstheorie und Verwandte Gebiete, 17, 168-180 (1971).

[5] D. Iglehart, Weak convergence in queueing theory, *Advances in Applied Probability*, 5, 570-594 (1973).

[6] J. Kingman, On queues in heavy traffic, J. R. Statist. Soc., 24, 383-392 (1962a).

[7] J. Kingman, The single server queue in heavy traffic, Proc. Camb. Phil. Soc., 57, 902-904(1962b).

[8] E. Kyprianou, The virtual waiting time of the $GI/G/1$ queue in heavy traffic, Advances in Applied Probability, 3, 249-268 (1971)

[9] S. Minkevičius, Weak convergence in multiphase queues, Lietuvos Matematikos Rinkinys, 26, 717-722 (1986) (in Russian).

[10] S. Minkevičius, Transient phenomena in multiphase queues, Lietuvos Matematikos Rinkinys, 31. 136-145 (1991).

[11] S. Minkevičius, On the law of the iterated logarithm in multiphase queues, Lietuvos Matematikos Rinkinys, 35, 360-366 (1995).

[12] S. Minkevičius, The probability limit theorem for the idle time of a customer in open queueing networks, Informatica (accepted, 2015).

[13] J.J. Morder, S. E. Elmaghraby (eds.). *Handbook of operational research models and applications*, Van Nostrand Reinhold, New York (1978).

[14] M. Reiman, Open queueing networks in heavy traffic, Mathematics of Operations Research, 9, 441-459 (1984).

[15] L. Sakalauskas, S. Minkevičius, On the law of the iterated logarithm in open queueing networks, European Journal of Operational Research, 120, 632-640 (2000).

[16] L. Takacs, Occupation time problems in the theory of queues. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, Heidelberg, New York, 98, 91-131 (1974).

[17] W. Whitt, Weak convergence theorems for priority queues: preemptive-resume discipline, Journal of Applied Probability, 8, 74=94 (1971).

[18] W. Whitt, Heavy traffic limit theorems for queues: a survey, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, New York, 98, 307-350 (1974).

Application of a Models Integration Module to the Cutting Slabs Problem in a Continuous Casting Machine

Konstantin Aksyonov, Anna Antonova, Elena Smoliy, Eugene Sysoletin, Alexei Sheklein
Dept. of Information Technology
Ural Federal University
Ekaterinburg, Russia
e-mail: wiper99@mail.ru, antonovaannas@gmail.com, bpsim.dss@gmail.com, unclesal@mail.ru, alexbrain@bk.ru

***Abstract*—This paper considers an application of a model integration module to the cutting slabs problem in a continuous casting machine. The model integration module is a part of the modeling subsystem of the metallurgical enterprise information system being developed. Developed information system is intended to decide the topical production problem: analysis and improvement the technological, logistical and organizational processes. This paper focuses on the interaction of the system modules and dives into the principles that the integration module is based on. Also, the development of a simulation model of cutting slabs in a continuous casting machine is considered with the use of the integration module. Simulation modeling is used to optimize the sequence of the melts supplied to the continuous casting machine. The goal of optimization is to reduce the deviation between the estimated and actual numbers of slabs, because such deviations can lead to missed deadlines. The integration module supports a multi agent simulation. Agents in the developed model are intended to describe the cutting slabs algorithm used by technologists in the metallurgical production. As a result of experiments with the model a solution has been found. The following sequence of melts should be supplied to the continuous casting machine: initially, the melts from the usual quality steel, followed by the melts from the high quality steel.**

***Keywords-model integration; simulation; agent-based modeling; continuous casting machine; automated system*.**

I. INTRODUCTION

Effectiveness of metallurgical production is tightly interconnected with perfection of technological, as well as logistical and business processes. The problems that can benefit from modeling are the following.

1. Industrial: a) irregularity of production load over a monthly period and other scheduling intervals; b) suboptimal load of production units of main technological process; c) as a result of the previous, suboptimal load of auxiliary and transportation equipment; d) unavailability of operative modification of the schedules and production graphs on technological reasons, related to the downtime of equipment on various reasons, interruption of stock supply.

2. Predictive modeling: a) analysis of effectiveness of suggested actions on new technological decisions; b) analysis of effectiveness of the innovative events on energy and resource saving; c) analysis of suggested actions on new business decisions at the level of processing facilities control, shop and inter-shop control and interaction; d) analysis of results of possible operative re-planning in case of unexpected failures, accidents and other emergencies.

Thus, creation and interaction of technological, logistical and business models is an important aspect for improvement of effectiveness of a metallurgical enterprise as a whole.

In addition, a feature of the problem of production processes execution analysis is the need to work with the processes model in real time. The processes model execution in real time means delivering the parameters from automated process control system (APCS) sensors to the model input and producing at the model output a set of control parameters transmitted in manufacturing execution system (MES-system). Real time is connected with the application of the control models for tracking and monitoring the current state of the processes. For example, in the cutting slabs problem there is a need to track in real time the signal supply for gas cutting for cutting the slab to withstand the specified dimensions of the slabs.

This paper focuses on the issue of the integration of enterprise information systems and intelligent automated system for tracking and monitoring the current state of the production process. We propose an approach for the integration of collection, storage and analysis of production processes, simulation of production processes in real time, and formation a recommendation for change in production processes. This approach is implemented in the models integration module of the metallurgical enterprise information system.

The remainder of the paper is organized as follows: Section 2 provides an overview of current state of the modeling tools. Section 3 presents a metallurgical enterprise information system architecture including systems modules interaction description. Section 4 introduces the application of the integration module to the decision of the cutting slabs problem in a continuous casting machine. Section 5 concludes this paper and explores future work.

II. CURRENT STATE OF TOOLS

In modern production enterprises, there is a problem of integration of heterogeneous information systems into a single information space. Such an integration requires a decision to be made for the following tasks:

- Synchronization of identifiers of the entities (processes, production units, etc.).

- Converting forms of information transfer. One accounting system can only work with the files, the other system allows communication via TCP/IP, the third system is a WEB-service.
- Converting formats of information transmitted. Attributes of the same production units are different in different information systems. Some attributes overlap (but have different names) and other attributes are specific to the information layer.
- Synchronization in time of incoming messages. Because the information systems are different (some records are maintained automatically, other involve human subjects), then, for example, we can have a message about creating a production unit arrive sooner than the message about the beginning of the process to create the production unit.

Today, there are several approaches to integrate heterogeneous information systems. The approach in which the information exchange takes place with the participation of the Enterprise Service Bus (ESB) [1] now. The basic idea of ESB is the message exchange between different IPs through a single point. But ESB is nothing more than a framework, based on which the mentioned problems will have to be solved. Not all information systems have ESB-adapters, so the transition from the form of data presentation in one system in the form of data representation in the other system usually has to be implemented within the ESB. In this paper, we propose an approach to the problem of integration of heterogeneous information systems aimed at solving the above mentioned problems, rather than to cover various protocols. The approach is implemented in the models integration module. This module allows using the integrated data from the different information systems in the real-time simulation.

The proposed approach has been verified on the data of metallurgical enterprise which has the following features: 1) the number of parameters of the life cycle of product unit is about 6-7 thousand units; 2) information on the parameters is stored in various information systems; 3) there are no adequate mathematical models of the processes. The modeling system in the proposed approach is used to describe and track in the real-time the dynamics of changes in the parameters of the production processes, optimization of production processes, and formation of recommendations for changes in the processes. Multi-agent simulation is used to formalize the expert knowledge of technologists about management of production processes.

The development trend of enterprise information systems focuses on wide application of Internet technologies. Currently the commercial modeling systems available on the market, including Plant Simulation [2], Simio [3], AnyLogic [4] are all desktop applications. Additional requirements for simulation modeling tools for team development of comprehensive simulation models include support for multi-user environment, and running simulation on the Internet. Comparison of Plant Simulation (PS), Simio (Sm), AnyLogic (AL) systems and modeling subsystem (MS) of the metallurgical enterprise information system is shown in Table 1.

TABLE I. ANALYSIS OF THE MODELING TOOLS

Comparison criteria	**PS**	**Sm**	**AL**	**MS**
Creation an enterprise processes model				
Description language of the technological, logistical and business processes	●	●	●	●
Creation of a multi agent model	○	●	●	●
Description of an agent knowledge base	○	○	○	●
Expert modeling	○	○	○	●
Optimization of enterprise processes				
Creation of an experiments plan	●	●	●	●
Changing the model parameters during the experiment	○	○	●	●
Use of heuristic methods to automatically search an optimal solution	●	●	●	●
Ability to use 2D / 3D animation	●	●	●	●
Formation recommendations on the elimination of bottlenecks	○	○	○	●
Execution model in real time	○	○	○	●
Convenient interface				
User interface (GUI / WEB)	GUI	GUI	GUI	WEB
Interface of a specialist in the subject area for model creation and process optimization	○	●	○	●

A metallurgical enterprise information system is developed in the Ural Federal University for the purpose of integrated application of the simulation and statistical analysis of production data for the problem of continuous improvement of the production process. The metallurgical enterprise information system is a web-oriented system for tracking, monitoring, modeling, analyzing and improving processes of steel products manufacturing [5]-[7].

Comparison of modeling tools showed that the most functionality is included in the modeling subsystem of the metallurgical enterprise information system and AnyLogic. Only modeling subsystem of the developed information system is focusing on the execution model in real time via web-oriented interface.

At the moment, the software-as-a-service (SaaS) technology is the most convenient in use, optimal in performance and client software requirements. The end user in this case is the analyst or decision making person. The modeling subsystem of the metallurgical enterprise information system includes a model integration module. This module uses a service oriented approach [5].

All of the analyzed modeling tools except MS cannot be included into the metallurgical enterprise information system because they do not support the real-time simulation and parameters exchange with the corporate information systems of the metallurgical enterprise.

III. METALLURGICAL ENTERPRISE INFORMATION SYSTEM ARCHITECTURE

The metallurgical enterprise information system consists of the following subsystems: subsystem of the data collection and analysis of production and modeling subsystem. The architecture of the metallurgical enterprise information system is shown in Figure 1.

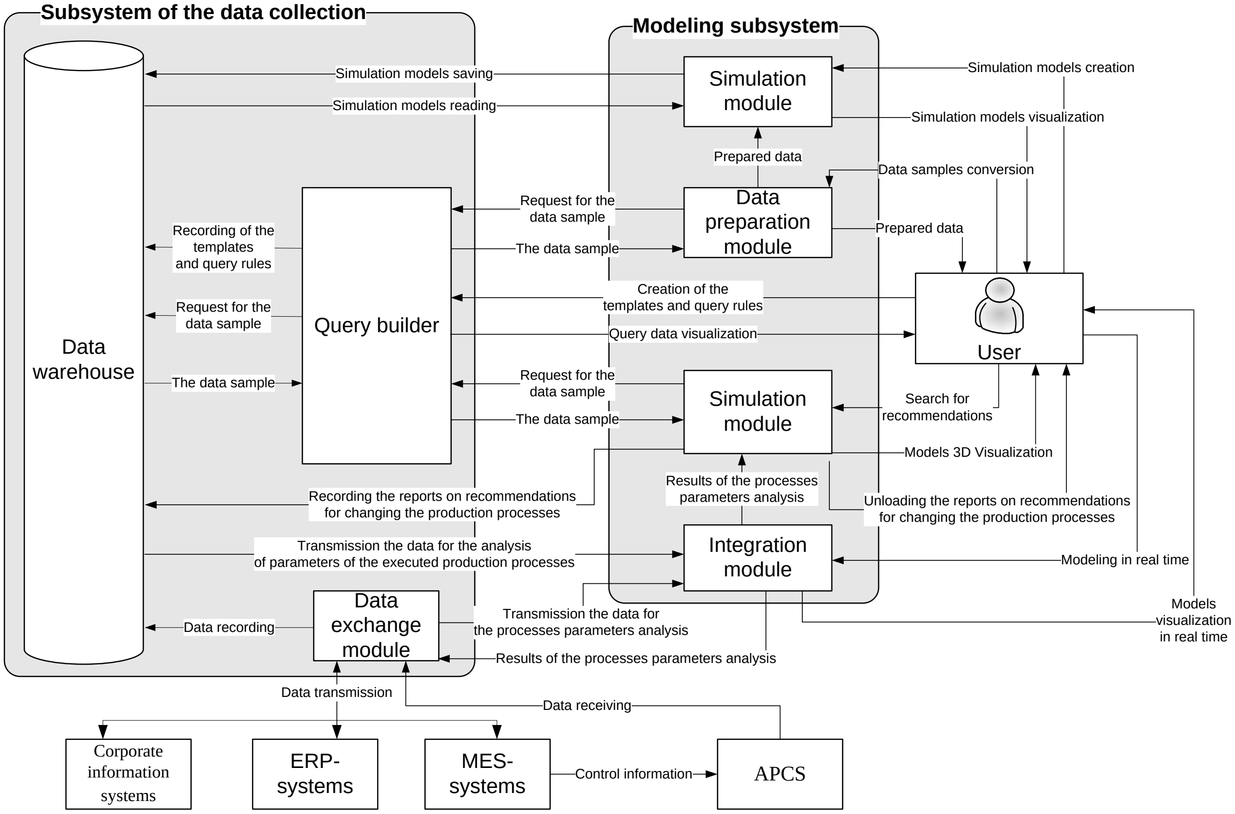

Subsystem of the data collection
Modeling subsystem
Data warehouse
Simulation models saving
Simulation models reading
Simulation module
Simulation models creation
Simulation models visualization
Prepared data
Data samples conversion
Data preparation module
Request for the data sample
The data sample
Prepared data
Recording of the templates and query rules
Query builder
Creation of the templates and query rules
Query data visualization
User
Request for the data sample
The data sample
Request for the data sample
The data sample
Simulation module
Search for recommendations
Models 3D Visualization
Results of the processes parameters analysis
Recording the reports on recommendations for changing the production processes
Unloading the reports on recommendations for changing the production processes
Transmission the data for the analysis of parameters of the executed production processes
Integration module
Modeling in real time
Models visualization in real time
Data recording
Data exchange module
Transmission the data for the processes parameters analysis
Results of the processes parameters analysis
Data transmission
Data receiving
Corporate information systems
ERP-systems
MES-systems
Control information
APCS

Figure 1. Architecture of the metallurgical enterprise information system

The subsystem of the data collection consists of the following modules: data exchange module, query builder [7], and data warehouse. The modeling subsystem consists of the following modules: data preparation module, simulation module, and integration module.

The data exchange module is intended for two-way communication between the modules of the developed system and enterprise information systems. The query builder module is intended to construct of requests for issuance of data on production processes without the involvement of information technology (IT) professionals.

The data warehouse module is intended to accumulate and store the information about production processes from the following data sources: corporate information systems, MES-systems, and APCS. The data preparation module is intended to transform and analyze the data from the data warehouse. The prepared data is used in the simulation module. The simulation module is intended to: 1) create and execute simulation models of the production processes with the use of 2D/3D animation, 2) receive data from the data preparation module, query builder, and integration module, 3) optimize the production processes, and 4) form recommendations on the elimination of bottlenecks.

A. Interaction of the Integration Module with the Data Exchange Module

The first option of interaction between the integration module and the module for data exchange with the enterprise automated system suggested the use of “Publisher – Subscriber” mechanism [8]. In this case, the integration module subscribed for a specific subject, and the data exchange module published all occurring events, distinguishing these by the subject. Thus, the integration module obtained only the required events. A significant disadvantage of this option is that the integration module has to reserve its own resources for the channel listening and awaiting the new information.

The main ideology of the system under discussion is the asynchrony of operation and the use of non-blocking operations of input-output. At this moment, the integration module uses the events mechanism, which is identical for all parts of the system. The event exchange is based on the Socket.IO protocol, on top of which the Java framework netty.io is implemented. Netty.io is the event oriented library for the implementation of network exchange using various protocols, including UDP, HTTP/HTTPS.

The data exchange module receives the emerging events on any of the listed protocols, providing the maximum versatility of its operation. At the same time, all events are produced using the only protocol, which is Socket.IO. Such approach allows maximum unification of event exchange, providing the feature of receiving messages both from Java based modules and from in-browser clients, written in JavaScript. Work through Socket.IO requires registration of callback procedures. Any waits in this case are eliminated.

Apart from this, the use of event mechanism conforms to the Model-View-Controller (MVC) concept [9], since the processes of receiving, possible data pre-processing, model execution and display of results are fully isolated. This allows multiple views of the same working model at the same time, e.g., the work may be represented in 2D and 3D, or the specific user may have the aggregated information on the shop or the full detail on a certain mechanism [5].

B. Data Exchange Mechanism Between the Modules

If model operation permits the real-time mode, the integration module subscribes for the necessary parameters. Subscription checks previously loaded models, which avoids double subscription for the same parameter (see Figure 2).

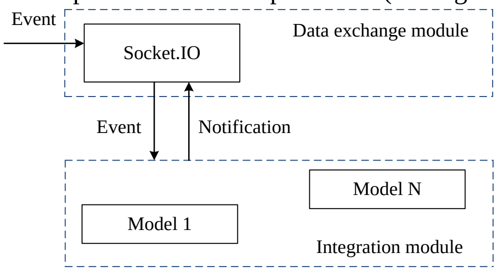


Figure 2. Receiving data from the data exchange module

The integration module launches the models and transfers the results of their operation into the data exchange module. In the first step, the fully loaded and executed test model modifies the state of its inputs. Next, the parameter is searched in the database. When the data exchange module receives the «Event_ParamValue_Got» event, it runs a number of checks. The received value must be of expected type (string, integer, etc.). The values themselves are written in the pre-defined form, depending on parameter kind (vector, scalar, linked list, synthetic).

Thus, before the «parameter value changed» event is received, the parameter itself needs to be stored in the database as an object with certain features. If this is not done, the processing of received value will cause an error.

The integration module interacts with the database in two cases: 1) when loading the model, since all information is stored in the database, 2) when model operation requires processing of statistical data, since this data is obtained beforehand and stored in the database. In the first case, the integration module receives a “Start” signal, which has the unique model identifier. Knowing it, the integration module accesses the database and selects the corresponding model. In the second case, the integration module receives the required parameter identifiers from the model as well as the desired time interval. Next, it selects the values of parameters from the database and forwards them to the model.

C. Multi-agent Architecture of the Integration Module

The multi-agent architecture of the integration module includes the following agents.

- Data exchange agent (used to update the model parameters and transfer the results of the experiment in the corporate information systems).
- Simulation agent (used to solve the process management problems in real time based on real-time models).
- Messaging agent (used for interaction between the data exchange agent and simulation agent).

The simulation agent allows to build models with the use of the notation of multi-agent resource conversion processes (MRCP) [6][10]. According to the MRCP notation, model's nodes are either agents or operations.

The simulation agent is based on the InteRRaP architecture [11] as the most appropriate for problem domain. In accordance with InteRRaP architecture common concept, a simulation agent model is represented in four levels (Figure 3).

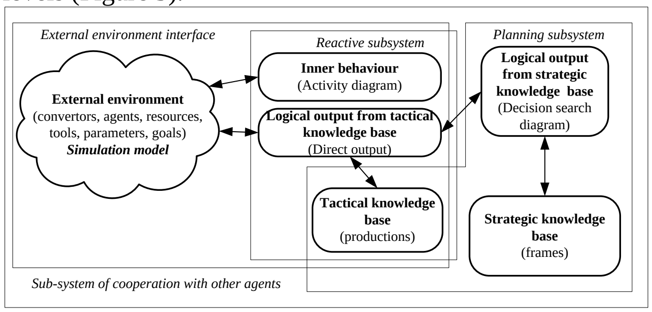

External environment interface
Reactive subsystem
Planning subsystem
Inner behaviour
(Activity diagram)
Logical output from strategic knowledge base
(Decision search diagram)
External environment
(convertors, agents, resources, tools, parameters, goals)
Simulation model
Logical output from tactical knowledge base
(Direct output)
Tactical knowledge base
(productions)
Strategic knowledge base
(frames)
Sub-system of cooperation with other agents

Figure 3. BPsim agent hybrid architecture

1. Sub-system of cooperation with other agents corresponds to the following MRCP elements: converters, resources, tools, parameters, goals.
2. External environment interface and reactive behavior components are implemented in form of agent rule’s base and inference machine (simulation algorithm).
3. Reactive subsystem performs the following actions: receives tasks from the external environment, places tasks in a goal stack, collates the goal stack in accordance with the adopted goal ranging strategy, selects a top goal from the stack, and searches on the knowledge base.
4. Local planning subsystem purpose is effective search for decisions in complex situations (e.g., when goal achievement requires several steps or several ways for goal achievement are available). Local planning component is based on a frame expert system. The frame-concept and conceptual-graph based approach are utilized for knowledge formalization.

IV. Application a Models Integration Module to the Cutting Slabs Problem

A. Statement of the Problem

In metallurgical production, special attention is paid to the improvement of the continuous casting process in order to increase the share of steel produced at continuous casting machine (CCM). The effectiveness of the CCM has a direct impact on the quality and cost of the production manufactured in the subsequent process stages. In the study of the physical and mechanical processes of continuous casting affecting the quality of the finished product, a method of mathematical modeling is widely spread [12]. In the study of the logistics and organizational (business) processes of continuous casting, a simulation method shows good results as applied to the optimization of production planning [13]. Development of simulation models of CCM technological processes affecting the quality of the finished product is topical.

We consider the development of a cutting slabs model with the use of the simulation module of the metallurgical enterprise information system. The developed model is run in the integration module, and the series of the melts are fed to the model input in real time (baseline experiment).

A casting ladle with melt goes to the CCM, where the liquid steel is cast; the ingot is cooled and cut into slabs. A CCM intermediate ladle distributes steel on the streams and allows continuous casting when replacing an empty casting ladle by full ladle. Metal is cooled in the water-cooled mould; the metal ingot is pulled in a secondary cooling zone (SCZ). For gas cutting (GC) work in automatic mode, technologists use an algorithm for cutting the ingot into slabs on the melts border.

When the signal "Start pouring of the melt from the casting ladle" is generated, the estimated length of the melt L^{i}_{est} is determined by the formula:

$$L^{i}_{est} = \frac{M^{i}}{2 \cdot \rho^{i} \cdot F} \qquad (1)$$

Formula (1) shows that the melt length is proportional to the steel mass with a coefficient depending on the steel density and cross-section of the finished slabs.

We consider the work of two-strand CCM with the following characteristics: the pulling speed of the ingot from the mold v is 0.8 meters per minute; the mold length is 1 meter, the SCZ length is 50 meters. A diagram of cutting the ingot of the melt (i-1) based on the desired slab length DSL^{i-1} and new melt start signal is shown in Figure 4. The parameter P is determined by measuring as shown in Figure 4.

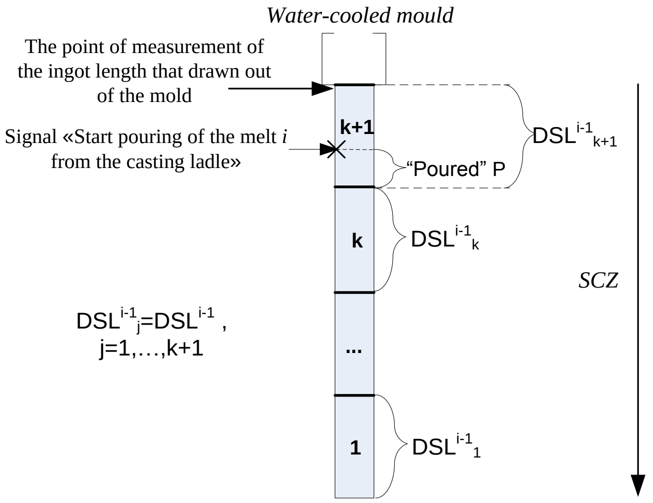

Water-cooled mould
The point of measurement of the ingot length that drawn out of the mold
Signal «Start pouring of the melt i from the casting ladle»
k+1
"Poured" P
DSL^{i-1}_{k+1}
k
DSL^{i-1}_k
SCZ
DSL^{i-1}_j=DSL^{i-1} , j=1,…,k+1
...
1
DSL^{i-1}_1

Figure 4. Diagram of cutting the ingot of the melt (i-1)

We analyze the CCM work on bottling series of 10 melts. Each melt i is characterized by the following parameters: steel weight in the casting ladle M^{i} (kg); steel grade (usual quality steel 'A' with density ρ^{i} equal to 7280 kg/m^3 and high quality steel 'B' with ρ^{i} equal to 7850 kg/m^3); desired slab length DSL^{i} (m); slab sectional area F equal to 0.225 m^2.

Table 2 shows the algorithm for cutting the ingot into slabs on the melts border. This algorithm is used by technologists in order to determine the actual number of slabs K^{i}_{act}. The estimated number of slabs K^{i}_{est} is defined as the ratio of the estimated melt length L^{i}_{est} to the desired slab length DSL^{i}.

We minimize the total number of positive and negative deviations in the slabs quantity S by changing the sequence of the supply of melts on the CCM. The total number of positive and negative deviations in the slabs quantity was determine by the formula:

$$S = 2 \cdot \sum_{i} \left| \Delta^{i}_{negative} \right| + \sum_{i} \Delta^{i}_{positive} \; , \qquad (2)$$

where $\Delta^{i}_{negative}$ – negative deviation in the slabs quantity of the melt i: $\Delta^{i}_{negative} = K^{i}_{act} - K^{i}_{est} < 0$,

$\Delta^{i}_{positive}$ – positive deviation in the slabs quantity of the melt i: $\Delta^{i}_{positive} = K^{i}_{act} - K^{i}_{est} \geq 0$.

Formula (2) shows that the total number of deviations in the slabs quantity is growing faster when the growth in the number of negative deviations is observed.

TABLE II. THE ALGORITHM FOR CUTTING THE INGOT INTO SLABS ON THE MELTS BORDER

Transition type on the melts border	**The condition on the parameter *P* "Poured"**	**Decision**
The melt "the worst for the best" (steel grade of the melt i is worse than steel grade of the melt i-1)	$P \leq 200cm$	In the melt (i-1) there are (k-1) slabs with a slab length DSL^{i-1}. Slab k with a new desired length DSL^{i} will be the first in the melt i.
	$P > 200cm$	In the melt (i-1) there are k slabs with a slab length DSL^{i-1}. Slab (k+1) with a new desired length DSL^{i} will be the first in the melt i.
The melt "equivalent " (steel grades of the melts i-1 and i are equal in quality)	$P = 0cm$	In the melt (i-1) there are (k-1) slabs with a slab length DSL^{i-1}. Slab k with a new desired length DSL^{i} will be the first in the melt i.
	$P > 0cm$	In the melt (i-1) there are k slabs with a slab length DSL^{i-1}. Slab (k+1) with a new desired length DSL^{i} will be the first in the melt i.
The melt "the best to the worst" (steel grade of the melt i is best than steel grade of the melt i-1)	$(DSL^{i-1}_{k+1} - P) \leq 600\,cm$	In the melt (i-1) there are (k+2) slabs with a slab length DSL^{i-1}. Slab (k+3) with a new desired length DSL^{i} will be the first in the melt i.
	$(DSL^{i-1}_{k+1} - P) > 600\,cm$	In the melt (i-1) there are (k+1) slabs with a slab length DSL^{i-1}. Slab (k+2) with a new desired length DSL^{i} will be the first in the melt i.

The negative deviation is such deviation, where the actual number of slabs is less than the estimated (expected) number of slabs. This situation is bad, because the order will not be satisfied. The number of positive deviations also affects the total number of deviations in the slabs quantity. The positive deviation is such deviation, where the actual number of slabs is more than the estimated (expected) number of slabs. This situation is not so good, because the additional slabs will not be paid.

B. Development of the Simulation Model of Cutting Slabs

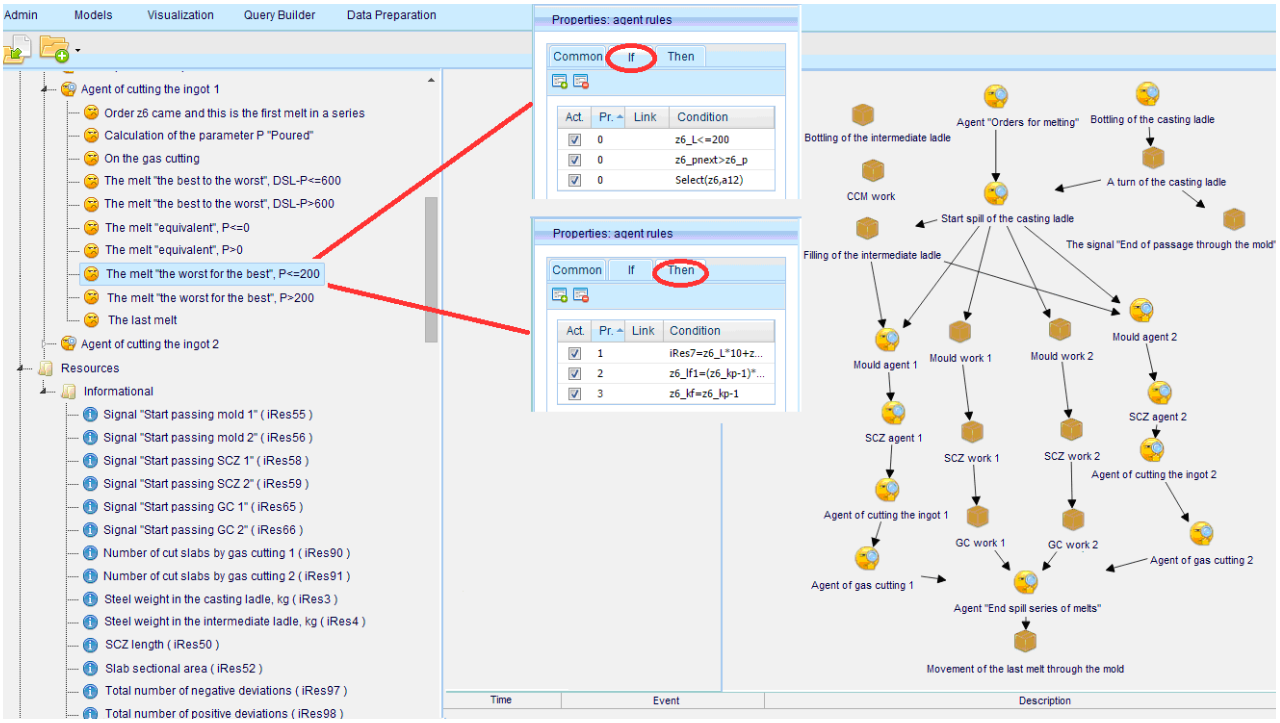
The structure of the simulation model developed via the simulation module of the metallurgical enterprise information system is shown in Figure 5 (right side). Agents in the model of cutting slabs are used to implement the logic to process the orders and to manage the order's attributes. Operations in the model are used to visualize the duration of the work of CCM elements for each stream: mold, secondary cooling zone and gas cutting. The model structure can be divided into three work units: 1) description of the casting ladle (intermediate ladle) state; 2) description of the CCM elements work; 3) description of the generation and removal of the orders. Seven orders have been described in the model. The order *z1* "Melt order" is the main order that collects data from all orders about temporal characteristics of the CCM elements work and the number of slabs for each melt. Orders *z2* and *z3* «Melting through the water-cooled mould 1" and "Melting through the water-cooled mould 2" are used to describe the logic of two streams of water-cooled moulds and then removed. Orders *z4* and *z5* «Melting through SCZ 1" and "Melting through SCZ 2" are used to describe the logic of two streams of the CCM secondary cooling and then removed. Orders z6 and z7 «Melting through GC 1" and "Melting through GC 2" are used to describe the logic of two streams of the gas cutting and then removed. Each order has its own attributes, for example, attribute *z6_pnext* contains the steel density for the next melt.

The algorithm for cutting the ingot into slabs on the melts border is described using agents of cutting the ingot for each CCM stream. A knowledge base of agent of cutting the ingot 1 (for stream 1) is shown in Figure 5 (in elements tree on left side). The agent knowledge base comprises situations described in Table 2. Separate situation is a rule of the form "If - Then" built using model variables (resources and orders *z6* attributes). The water-cooled mould agent knowledge base contains "If - Then" rules to ensure the work with the orders *z2*/*z3* attributes and generation of the orders *z4*/*z5*, which is transferred to the SCZ agent. The SCZ agent knowledge base contains "If - Then" rules to ensure the work with the orders *z4*/*z5* attributes and generation of the orders *z6*/*z7*, which is transferred to the ingot cutting agent.

C. Experiments Results Analysis

We consider the experiments with the developed model in the integration module of the modeling subsystem. Table 3 shows the simulation results for the initial input data. The following value of the output characteristic has been obtained: total number of positive and negative deviations in the slabs quantity *S*=4.

We conducted and analyzed a series of seven experiments on the supply chains of the melts with different interleaving DSL parameters and steel grade. The results of the experiments are shown in Figure 6 as a distribution of the total number of deviations in the slabs quantity *S* depending on the experiment.


Admin
Models
Visualization
Query Builder
Data Preparation
Properties: agent rules
Common
If
Then
Properties: agent rules
Common
If
Then
Bottling of the intermediate ladle
Agent "Orders for melting"
Bottling of the casting ladle
CCM work
A turn of the casting ladle
Start spill of the casting ladle
Filling of the intermediate ladle
The signal "End of passage through the mold"
Mould agent 1
Mould work 1
Mould work 2
Mould agent 2
SCZ agent 1
SCZ work 1
SCZ work 2
SCZ agent 2
Agent of cutting the ingot 2
Agent of cutting the ingot 1
GC work 1
GC work 2
Agent of gas cutting 2
Agent of gas cutting 1
Agent "End spill series of melts"
Movement of the last melt through the mold
Time
Event
Description

- Agent of cutting the ingot 1
  - Order z6 came and this is the first melt in a series
  - Calculation of the parameter P "Poured"
  - On the gas cutting
  - The melt "the best to the worst", DSL-P<=600
  - The melt "the best to the worst", DSL-P>600
  - The melt "equivalent", P<=0
  - The melt "equivalent", P>0
  - The melt "the worst for the best", P<=200
  - The melt "the worst for the best", P>200
  - The last melt
- Agent of cutting the ingot 2
- Resources
  - Informational
    - Signal "Start passing mold 1" (iRes55)
    - Signal "Start passing mold 2" (iRes56)
    - Signal "Start passing SCZ 1" (iRes58)
    - Signal "Start passing SCZ 2" (iRes59)
    - Signal "Start passing GC 1" (iRes65)
    - Signal "Start passing GC 2" (iRes66)
    - Number of cut slabs by gas cutting 1 (iRes90)
    - Number of cut slabs by gas cutting 2 (iRes91)
    - Steel weight in the casting ladle, kg (iRes3)
    - Steel weight in the intermediate ladle, kg (iRes4)
    - SCZ length (iRes50)
    - Slab sectional area (iRes52)
    - Total number of negative deviations (iRes97)
    - Total number of positive deviations (iRes98)

Act.	Pr.	Link	Condition
☑	0		z6_L<=200
☑	0		z6_pnext>z6_p
☑	0		Select(z6,a12)

Act.	Pr.	Link	Condition
☑	1		iRes7=z6_L*10+z...
☑	2		z6_lf1=(z6_kp-1)*...
☑	3		z6_kf=z6_kp-1

Figure 5. Structure of the simulation model developed in the modeling subsystem of the metallurgical enterprise information system

TABLE III. BASELINE EXPERIMENT RESULTS

Input model parameters			Output model parameters			
№ of the melt, i	***Steel grade type***	***DSL^i (mm)***	***Actual length of the melt, L^i_{act} (mm)***	***Actual number of slabs, K^i_{act}***	***Estimated number of slabs, K^i_{est}***	***Number of deviations in the slabs quantity, Δ^i***
1	B	5500	39039	7	7	0
2	B	7900	48066	6	5	1
3	A	12000	36504	3	3	0
4	A	9100	46135	5	6	-1
5	B	6000	48672	9	9	0
6	B	8800	53538	6	5	1
7	A	6700	40764	5	5	0
8	A	7300	44412	6	6	0
9	B	10100	51205	5	5	0
10	B	11600	35286	3	3	0
Total number of positive and negative deviations in the slabs quantity, S						4

As Figure 6 shows, experiments with the best results are experiments №2 and №5: they are characterized by the absence of negative deviations in the number of slabs. In these experiments, the following sequence of the melts has been supplied on the CCM: initially, the melts from the usual quality steel, followed the melts from the high quality steel.

The impact of ascending / descending DSL on the number of deviations in the slabs quantity has not been identified.

In this section, a simulation model of cutting slabs in the continuous casting machine has been described using a modeling subsystem of the metallurgical enterprise information system. The developed model uses agent-based modeling in order to represent the knowledge of technologists.

The model of cutting slabs has been used in the integration module to solve the problem of optimization the CCM melts sequence.

V. CONCLUSION AND FUTURE WORK

This paper introduces the perspectives of simulation modeling in metallurgical production. The architecture of the developed metallurgical enterprise information system has been described. Among all of the system modules the integration module has been highlighted as a module that launches the simulation models in real time. During real-time modeling, the input parameters are fed to the model from the corporate information systems, MES-systems, or APCS, and the output parameters formed by the model are translated to the data exchange module. This paper focuses on the description of the systems modules interaction and describes the integration module work principles.

The integration module has been applied to the decision of the cutting slabs problem in the continuous casting machine. As a result of the experiments, the following recommendations have been obtained to optimize the CCM processes: it is necessary to supply the melts to CCM in the following sequence – initially, the melts from the usual quality steel, followed the melts from the high quality steel. These results are consistent with those obtained on the production. Use of simulation modeling for analysis of technological, logistical and business problems of an enterprise is a perspective direction. The method of simulation models integration that has been implemented in the developed system has successfully passed the test.

The aim of future research is to apply the query builder and the data preparation modules on the stage of receiving input modeling data from real production.

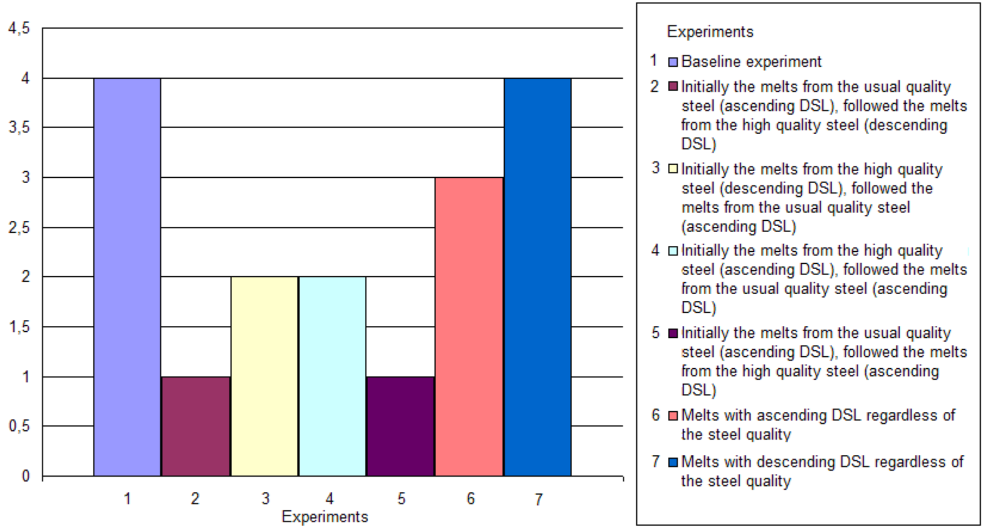


Figure 6. Distribution of the total number of deviations in the slabs quantity *S* depending on the experiment

ACKNOWLEDGMENT

This research has been conducted under the terms of contract № 02.G25.31.0055 (project 2012-218-03-167) with financial support from the Ministry of Education and Science of the Russian Federation.

REFERENCES

[1] D. Chappell, Enterprise Service Bus, O'Reilly, 2004.

[2] Plant Simulation. The official web site, Available from: http://www.plm.automation.siemens.com/en_us/products/tecnomatix/plant_design/plant_simulation.shtml. [Retrieved: May 2015]

[3] Modeling system Simio. The official web site, Available from: http://www.simio.com. [Retrieved: May 2015]

[4] Modeling system AnyLogic. The official web site, Available from: http://www.anylogic.com. [Retrieved: May 2015]

[5] K. A. Aksyonov, E. A. Bykov, O. P. Aksyonova, and A. S. Antonova, "Development of real-time simulation models: integration with enterprise information systems," Proceedings of ICCGI 2014: The Ninth International Multi-Conference on Computing in the Global Information Technology, 22-26 June 2014, Sevilla, pp. 45–50.

[6] K. A. Aksyonov, I. A. Spitsina, E. G. Sysoletin, O. P. Aksyonova, and E. F. Smoliy, "Multi-agent approach for the metallurgical enterprise information system development," Proceedings of 24th Int. Crimean Conference CriMiCo 2014: Microwave & Telecommunication Technology, 7-13 September 2014, Sevastopol, vol. 1, pp. 437–438.

[7] A. Borodin, Y. Kiselev, S. Mirvoda, and S. Porshnev, "On design of domain-specific query language for the metallurgical industry," Proceedings of 11th Int. Conference BDAS 2015: Beyond Databases, Architectures and Structures: Communications in Computer and Information Science, 26-29 May 2015, Ustron, Poland, vol. 521, pp. 505–515.

[8] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec, "The many faces of publish/subscribe," ACM Computing Surveys, vol. 35, no. 2, 2003, pp. 114–131.

[9] T. M. H. Reenskaug, Working with objects: the OOram software engineering method. TASKON, 1995.

[10] K. A. Aksyonov, Theory and practice of decision support tools. Germany, Saarbrucken: LAP LAMBERT Academic Publishing GmbH & Co. KG, 2011.

[11] J. P. Muller and M. Pischel, The agent architecture InteRRap: concept and application, German Research Center for Artificial Intelligence (DFKI), 1993.

[12] M. Barna, M. Javurek, J. Reiter, J. Watzinger, B. Kaufmann, and M. Kirschen, "Numerical simulations of the continuous casting of steel with electromagnetic braking and stirring," International Journal of Multiphysics, Series Special Edition, 2011, pp. 231–238.

[13] AnyLogic: case study, Available from: http://www.anylogic.com/case-studies/chelyabinsk-metallurgical-plant-uses-a-simulation-model-electric-furnace-melting-shop. [Retrieved: May 2015]

An Idea On Infinite Horizon Decision Support For Rule-based Process Models

Michaela Baumann*, Michael Heinrich Baumann†, and Stefan Jablonski*
*Institute for Computer Science
†Institute for Mathematics
University of Bayreuth, Germany
Email: {michaela.baumann,michael.baumann,stefan.jablonski}@uni-bayreuth.de

Abstract—**In recent years, process models tend to turn away from common procedural models to more flexible, rule-based models. The models are characterized by the fact that in each execution step users usually have to decide between several rule-consistent tasks to perform next. Precise execution paths are not given, which is why adequate execution support needs to be provided. Simulation is one means to facilitate the users' decisions. In this context, we suggest an execution simulation tool with an infinite horizon, i.e., in each (simulated) step, users are informed about the tasks that in any case still need to be done to properly finish one process instance, and about tasks that may no longer be executed. The forecasts consider an actual or a simulated history of the process instance and the rules given by the model.**

Keywords–Process execution; Rule-based process models; Process decision support;

I. Introduction

In many fields of economy, industry, and research, process models are used for supporting the execution of operating processes, for designing work steps, for documentation purposes, etc. Usually, these process models are a sort of procedural process models, where the execution order of the process steps is prescribed through the control flow. Other execution orders than the prescribed ones are not provided. This is why computational offloading ("the extent to which differential external representations reduce the amount of cognitive effort required to solve informationally equivalent problems" [1]) is quite well achieved in procedural process models. For rule-based process models, this is not the case [2], as they take a different modeling and representation approach. They are typically used when procedural process models are too restrictive or get too complicated when complex facts shall be displayed. The approach of rule-based process models is to provide a set of tasks, firstly without stating any execution order, and then to restrict all possible execution orders by adding rules or constraints that should be met during the execution. An example for such a rule could be: "If task A has been executed, afterwards task C needs to be eventually executed, too". Thus, especially for rule-based process models, guidance for the user through the process is necessary, as the execution sequences leading to a proper process completion are not easy to see [3].

In this paper, we do not want to answer the question of which tasks may be executed in the next step, with a certain process history underlying. This has been done in other work, e.g., in [4] for ConDec models via automata, and is not part of the work at hand. Tasks for the next step have to be chosen in a way that every resulting process history is model conform and that dead ends are avoided. Furthermore, we need process models that do not contain conflicting constraints [4].

For run-time support, recommendations for effective execution [5] can be given. However, these recommendations are usually based on past experiences and need a specific goal, i.e., a rating of experiences in terms of desirability [6], as input. Parts of the executable tasks are hidden from the executing agent, i.e., a preselection has occurred. The decision support we head for is somehow different, as we do not intend to give recommendations based on a specific goal (as input into the system) but to provide the agent an overview over *the impact of each of his decisions*. He can then decide, according to the overview and a goal (which is only in his mind), which step to execute next. The system and the model do not need to be changed, which may cause history-based violations when done at run-time [6]. The questions that shall be answered by the support are the following: "Which tasks still need to be executed during the process instance?", "Which tasks may/can still be executed eventually during the process instance?", "What changes apply to the answers of the two preceding questions if one (or more) certain task is executed next?" As one can see, there is no limit of steps till the end of an instance for answering these questions, which is why we talk of *infinite horizon* in this context. A use case for this approach could be the following example situation: An employee has noticed that his colleague is overloaded with work, and thus he wants to finish the process without involving this colleague, i.e., avoid certain tasks, if possible.

The work proceeds as follows: Section II proposes the infinite horizon decision support with help of examples, Section III concludes with some features of the approach, remaining questions that still need to be answered, and suggestions for future work.

II. Idea: Infinite Horizon Decision Support

We want to present our approach with a short example. Therefore, we consider four rules: the $existence$ rule, the $response$ rule, the $precedence$ rule, and the $chainResponse$ rule. They are defined as follows:

i) $existence(A, m, n)$: Task A must at least be executed m times and may at most be executed n times ($m \leq n$)
ii) $response(A, B)$: If task A appears in the process instance, then task B has to appear after A, too
iii) $precedence(A, B)$: Task B can only be executed if task A has already been executed, i.e., already appears in the process history
iv) $chainResponse(A, B)$: Every execution of task A has to be directly followed by B

$\forall A \in \mathcal{A}$: $man(A) \leftarrow 0$; $opt(A) \leftarrow \infty$;

Figure 1: Initialization of mandatory and optional values

$\forall\ existence(A, m, n) \in \mathcal{R}$:
If $start$: $man(A) \leftarrow m$; If $start$: $opt(A) \leftarrow n$;

Figure 2: Update rules for process rule $existence$

For every task in the process model two states are recorded: *mandatory* and *optional*. The initialization of these states is conform to the rule-based approach. Let $\mathcal{A}$ denote the set of all tasks and $\mathcal{R}$ the set of all rules. At first, without considering any rules, no task must be executed ($\forall A \in \mathcal{A}: \ man(A) = 0$) but may be executed arbitrarily often ($\forall A : \ opt(A) = \infty$). The process history is stored in variable h. At the beginning, the history is empty: $h = \odot$. The task executed in the previous step is given by $\ell(h) \in \mathcal{A} \cup \{\mathtt{NA}\}$.

After initialization of the status values for all tasks (Figure 1), the values are sequentially restricted according to the update rules (Figures 2–5), where function $start$ denotes the beginning of a process instance and $exec(\cdot)$ the execution of a task. After each task execution, both status values of the corresponding task are reduced by 1, if possible, before considering the update rules and adjusting the status values according to them. The list of all rules is processed sequentially, as many times, until in one run nothing more changes.

The update rules can be divided into three different kinds of rules. One type are the start and execution rules (If $start$, If $exec(\cdot)$). The start rules only need to be processed once after starting the process and can be skipped for the rest of the process after the first task execution. The execution rules need to be processed once after each task execution and can be skipped at the beginning. The second type are the indirect status update rules (all other rules in the example Figures 2–5 except for the last one in Figure 5), triggered through chain reactions caused by start and execution rules. Rules of the third type need to hold permanently and have no special trigger constraint, like the last rule in Figure 5 or the rule $opt(A) \geq man(A)$. The reason for the last update rule (resulting from $chainResponse$) in Figure 5 is: As after A, task B must always follow directly, then B needs to be done at least as many times as A (plus 1, if A was the most recently task). If B needs to be done more often anyway, then nothing changes.

A possible prototype could look like the design draft in Figure 6, where two situations are shown. After having executed tasks A, A, C, F, and B, the tasks that are executable next in the first situation are tasks A and C. Note, that the update rules do not derive these next-executable tasks (that one with horizon step $n = 1$). The update rules rather determine the column on the right ($n = \infty$), which says that there exist possible execution paths for each task A to F, and that tasks B and C need to be executed in all of these paths to successfully finish the process execution. In the second situation, the history is still the same, but it is simulated how the infinite horizon changes if task C would be executed next (if the history was A.A.C.F.B.C). Now, task D may no longer be executed, no matter which task is chosen next (B, C, or E), and the status of C changes from mandatory to optional, so, $man(C) = 0$ and $opt(C) > 0$. This information, especially the infinite horizon after the simulation step (simulated execution of C), may help the agent to decide what to do next. The decision support can be expanded by using past execution histories to provide the agent information about average execution time for each step or about success rate of certain histories [5].

$\forall\ response(A, B) \in \mathcal{R}$:
If $exec(A)$: $man(B) \leftarrow \max\{man(B), 1\}$;
If $man(A) > 0$: $man(B) \leftarrow \max\{man(B), 1\}$;
If $opt(B) == 0$: $opt(A) \leftarrow 0$;

Figure 3: Update rules for process rule $response$

$\forall\ precedence(A, B) \in \mathcal{R}$:
If $opt(A) == 0 \wedge A \notin h$: $opt(B) \leftarrow 0$;
If $man(B) > 0 \wedge A \notin h$: $man(A) \leftarrow \min\{man(A), 1\}$;

Figure 4: Update rules for process rule $precedence$

$\forall\ chainResponse(A, B) \in \mathcal{R}$:
If $exec(A)$: $man(B) \leftarrow \max\{man(B), 1\}$;
If $opt(B) \neq \infty$:
  $opt(A) \leftarrow \min\{opt(B), opt(A) - \mathbb{1}_{\ell(h)==A}\}$;
$man(B) \leftarrow \max\{man(B), man(A) + \mathbb{1}_{\ell(h)==A}\}$;

Figure 5: Update rules for process rule $chainResponse$

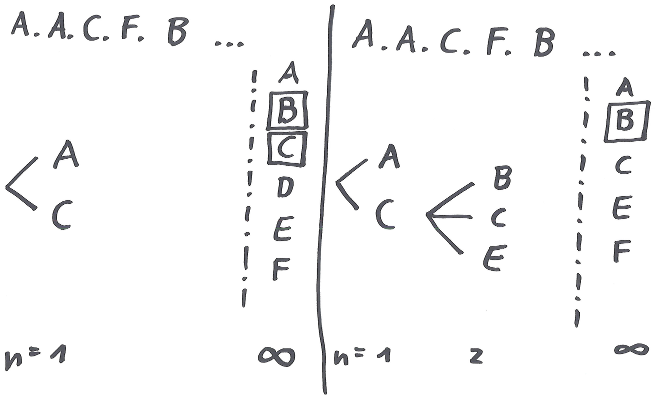


Figure 6: Prototypical design for an infinite horizon decision support tool

If the underlying process model is changed, then the set of update rules needs to be changed, too. The update rules corresponding to removed process rules, or even removed tasks, have to be eliminated, whereas new update rules, caused by added process rules, are included. For running process instances, there may occur two situations [7]: The current history is conform to the new set of process rules, then the new status values can be achieved by simulating their evolution according to the new set of update rules. If a so-called “history violation” [7] occurs, then we refer to [7] for handling the problem.

III. Features, Remaining Questions, and Future Work

The idea paper suggests a possibility for decision support for declarative process models. This decision support applies to an infinitely long forecast horizon. It makes use of the process rules and their implications to the states (mandatory and optional) of the tasks they refer to. At the moment, the focus lies only on rules concerning control-flow. It should be investigated if and how an extension to other process perspectives, like data and agents, is possible. Furthermore, instead of regarding tasks as single points in time, i.e., only their final execution is registered, it would be beneficial to split one task into different events, at least start and end. Nesting of tasks (subprocesses) could also be analyzed.

Repeatability and optionality of tasks [8] may be read off the status values, as well as dead activities when regarding the status values with empty history $h = \odot$. If one task A has $opt(A) = 0$ at the beginning after the first evaluation round of the update rules, it can never be executed. The question arises if it is possible to modify update rules so that conflicts can be detected. Also, the completeness of the list of update rules has to be proven.

A further issue would be to check, if the status values and update rules can be utilized for determining tasks that can be executed next (the part of the tool, that is assumed to be given at the moment). Perhaps this can be achieved through a checking like this: “If task A is executed next, then the constraint $opt(B) \geq man(B)$ (for an arbitrary task B) is violated”. Thus, A cannot be suggested now (in the next step) for execution. Automata like in [4] would not be needed in that case.

In the context of log-based recommendations, it could also be interesting to include reviews into the decision support system. Users could rate their decisions at some time after their execution which is valuable information in future. To improve the performance, once calculated status values could be stored as tables together with the respective history in a hash-based repository.

Acknowledgement

The work of Michael Heinrich Baumann is supported by a scholarship of “Hanns-Seidel-Stiftung (HSS)” which is funded by “Bundesministerium für Bildung und Forschung (BMBF)”. The authors want to thank Lars Ackermann and Stefan Schönig, both with University of Bayreuth.

References

[1] M. Scaife and Y. Rogers, “External cognition: how do graphical representations work?” *International Journal of Human-Computer Studies*, vol. 45, no. 2, pp. 185–213, 1996.

[2] S. Zugal, J. Pinggera, and B. Weber, “Creating declarative process models using test driven modeling suite,” in *IS Olympics: Information Systems in a Diverse World*, ser. LNBIP, S. Nurcan, Ed. Springer Berlin Heidelberg, 2012, vol. 107, pp. 16–32.

[3] W. M. van der Aalst, M. Weske, and D. Grnbauer, “Case handling: a new paradigm for business process support,” *Data & Knowledge Engineering*, vol. 53, no. 2, pp. 129 – 162, 2005.

[4] M. Pesic, “Contraint-based workflow management systems: Shifting control to users,” Ph.D. dissertation, Technische Universiteit Eindhoven, 2008.

[5] W. van der Aalst, M. Pesic, and H. Schonenberg, “Declarative workflows: Balancing between flexibility and support,” *Computer Science - Research and Development*, vol. 23, no. 2, pp. 99–113, 2009.

[6] M. Pesic, H. Schonenberg, and W. van der Aalst, “DECLARE: Full support for loosely-structured processes,” in *Enterprise Distributed Object Computing Conference, 2007. EDOC 2007. 11th IEEE International*, 2007, pp. 287–287.

[7] M. Pesic, M. Schonenberg, N. Sidorova, and W. van der Aalst, “Constraint-based workflow models: Change made easy,” in *On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS*, ser. LNCS, R. Meersman and Z. Tari, Eds. Springer Berlin Heidelberg, 2007, vol. 4803, pp. 77–94.

[8] M. Baumann, M. H. Baumann, and S. Jablonski, “On behavioral process model similarity matching: A centroid-based approach,” 2015, preprint. [Online]. Available: https://epub.uni-bayreuth.de/id/eprint/2051 [accessed 2015-07-18]

Integrating Crime Data by the Use of Generic Data Models

Dirk Frosch-Wilke
Institute of Business Information Systems
University of Applied Sciences Kiel
Kiel, Germany
e-mail: dirk.frosch-wilke@fh-kiel.de

Lennard Scheffler
Graduate of
University of Applied Sciences Kiel
Kiel, Germany
e-mail: l.scheffler@online.de

Abstract. Effective and efficient crime data analysis can facilitate police work and could increase success rates in crime fighting and crime prevention (e.g. predictive policing). Relevant crime data are usually stored in various heterogeneous systems within different police organizations and other federal agencies. For this reasons many crime fighting agencies are eager to build Data Warehouse Systems for integrating crime data and providing a single data source for crime data analysis and crime data mining. But these Data Warehouse projects are often faced with data quality problems due to incomplete integration of relevant crime data. The reason for this is often the variety, variability, and granularity of crime data in different IT-systems and the inflexibility of the data models used in Data Warehouses to handle these complex data characteristics. In this paper we apply concepts of Generic Data Modeling and Data Model Patterns in order to build data models for crime data which allows complete and consistent integration of crime data in Data Warehouses.

Keywords-Relational Data Modeling; Data Warehouse; Generic Data Modeling; Police Data, Data Model Pattern

I. Introduction

The research about Business Intelligence and Data Warehousing focused mainly on business [1]. However, the knowledge about how to integrate and analyze data is portable to other contexts. Especially the public sector that stores and manages many different data has now taken advantage of Business Intelligence concepts [2], [3], [4]. This work deals with how to provide high quality data models for the police context. Even though many police departments have noticed the benefit of data analysis, they often suffer from insufficient data integration and analysis options. Though the reasons for this are sometimes project specific, a more general reason is the quality of the data models especially in the crime fighting sector. Therefore this paper particularly deals with how to develop flexible and solid data models in order to integrate crime data from various sources on the long run.

The paper is organized as follows. In Section II we shortly describe the Application Domain. Section III provides the theoretical background for data model patterns as well as for Generic Data Modelling and thereby describing the Knowledge Base of our research.

In Section IV we present some of our design artifacts for high quality crime data models as well as the evaluation of these artifacts. The paper ends with a conclusion and prospects regarding further research activities.

II. Application Domain

Even though this research results in abstracted data model solutions that can be applied to various projects and contexts regarding to the police domain, the starting point has been a data warehouse project of some federal police department in Germany. Since the year 2012 both, the Business Intelligence Competence Center (BICC) as well as the Business Intelligence department of the federal IT-service provider maintain a data warehouse system in order to support strategic and operational crime fighting activities. This system is based on Inmon’s enterprise architecture [1] running on an Oracle database system. It integrates various operational data sources into a relational core area and provides different multi-dimensional data marts for analysis purposes. Outcomes can be visualized ether in charts, crosstabs or geographical maps. This system especially in terms of standard reporting is used by police officers as well as management and controlling.

The challenge of this project is to build flexible and consistent data models, which allow to easily integrating existing data sets as well as new kinds of data. The crime-fighting sector stands out due to a huge amount of different data. This means, documenting a criminal case brings up varying entities and objects depending on the particular circumstances. Thus, examining a single case, the amount of data is not down to a large number of similar objects but too many objects of different classes, e.g. persons, things, locations, actions and so on. For this reason, detailed relationships between objects usually can add most of the value for data analysis. Only inquiring objects of one class without considering relationships to other entities will not lead to satisfying results. Summarized, a data model for integrating crime data has to care about both, providing well-structured objects as well as detailed relationships that can handle most of possible object correlations.

Although this is a challenging task on its own, the developer has to care also about maintaining the data model in a flexible way in order to fit future requirements and simultaneously lower the complexity towards the end-user. Because of different reasons the mentioned project sticks to multi-dimensional modeling of data marts, where it can be difficult to transform the complex relational core model into a valid and high-performance multi-dimensional model. In particular, the many n:m-relationships can affect the size and complexity of fact tables enormously. Lastly, it is not possible to develop data models (neither relational nor multi-dimension) without profound knowledge of the problem domain.

III. Knowledge Base

A. Data Model Patterns

Back in the beginnings of software engineering in the sixties of the last century software development was understood as an artistic task [5]. Friedrich L. Bauer was the first to introduce the idea of a software developer as an engineer in 1969 [6]. Since then standardization in methods, processes, roles and techniques increased by following the example of conventional engineering. One of the most influential advantages has been the development of reusable parts, so called *patterns*, basically invented by Christopher Alexander [7]. While the use of patterns has already been known for many years in programming [8] it is now assigned to the area of conceptual data modeling e.g. by David C. Hay [9] or Len Silverston [10] and has been applied e.g. for dimensional modeling to the Data Warehouse context [11].

In the context of data modeling a pattern is understood as a template or guidance for developing parts of a data model in order to store data of a certain domain. Every pattern is abstracted to be reusable for a class of similar problems. Moreover, a pattern consists of three parts. The *context* defines in which condition the pattern can be used. The *problem* describes the challenge that has to be solved. Lastly there is an *abstract solution*, which has to be concretized by the user [12]. The probably greatest advantage in using data model patterns is to solve a problem once and then apply this solution to multiple data models. This leads to a better understanding of data models and decreases development costs. Furthermore, high quality patterns can highly increase the strength and flexibility of data models.

B. Generic Data Modeling

The benefits of patterns can increase data model quality, if the pattern itself is of high quality. To create significant data model patterns, there is a need for methodic definition. In 2007, Stephan Schneider introduced the concept of *Generic Data Modelling* as methodological background [13].

Many developers presuppose an adequate grade of their data models, which results in insufficient quality management. This approach often tempts to cover weaknesses of data models, which ends in low quality results. On top of that, these vulnerabilities are frequently not noticed before trying to embed new requirements. Therefore, the method of generic data modeling integrates the quality management into the process of modeling. Basically, a developer has to follow eleven rules exceeding known strategies like normalization (c.f. table 1). Following these rules is one important activity in achieving high quality data models. But it is also necessary to separate the concepts of *nature* and *roles*, which many data modeling professionals are lacking [13].

TABLE 1 RULES OF GENERIC DATA MODELING [13, p. 395 ff.]

Rule	Description
completeness	The data model represents the whole problem domain or the necessary parts.
correctness	All data model items are consistent and do not need to be interpreted.
minimality	The model is free of redundancy; no part could be removed without losing information content.
understandability	The model is significant to the reader by using meaningful naming for entities, properties and relationships.
simplicity	The data model is not unnecessarily complex.
flexibility	The data model enables easy integration of changed requirements or conditions.
stability	The data model offers ways to implement changed requirements or conditions without changing structure.
modularity	The data model is based on consistent model parts.
reusability	Model parts can be adopted directly or slightly modified to other model parts.
ability for integration	Data model parts can be combined to a complete model.
ability for implementation	The data model can be implemented on a database system.

A simple example may explain both concepts: In the problem domain of crime fighting the entities suspect, witness, and victim are obvious. A quick and even common data modeling solution would be to model these three as separate entities. But of course these identified entities have many common attributes like name, address, date of birth and so on. For this reason the abstraction of parts of each of these entities to a generalized entity named *person* is possible. *Person* reflects the *nature* of the former mentioned entities. Understanding these entities as *roles* of a *person* would complete the data model and make it even closer to reality (cf. Fig. 1). Moreover, separating between nature and role can help avoiding *hidden redundancies*. This becomes clear while looking at a suspect for instance, who is witness in the same record. By using the role concept, it is possible to store the key attributes of the person only once and then relate it to its roles. The other way would expect to store the person for each of its roles, which will expire in data inconsistencies.

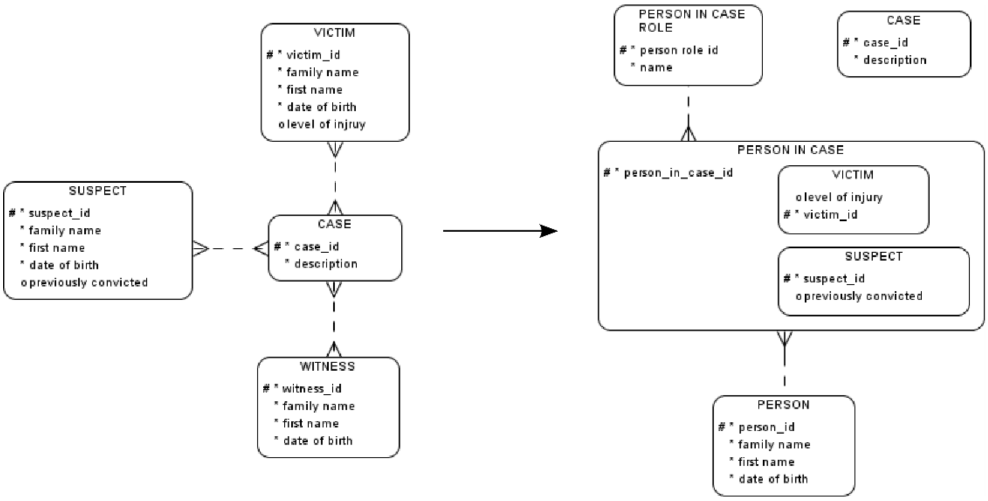


Figure 1. Abstraction by example

IV. Design Artifacts

The following part of this article explains some of our developed patterns for crime data integration within the above mentioned project (c.f. Section II). The presented data models are not necessarily complete but they build the fundamental architecture which allows an easy extension of the models by following the explained rules.

A. Characteristics

One of the main challenges in data modelling for crime fighting is the complexity of police work. The key to effective crime fighting is to store as many facts as possible that might be important for solving crime cases. In consequence, crime data is characterized by a very high level of granularity and heterogeneity in order to fulfill this requirement.

The initial data modeling process cannot account for every possible fact that might be relevant later on. Therefore, the data model has to fit abstraction requirements in order to store every relevant data in future. Otherwise, the limitation of storable attributes for describing a suspect for example could end up in bad analytical and crime fighting results because of information loss.

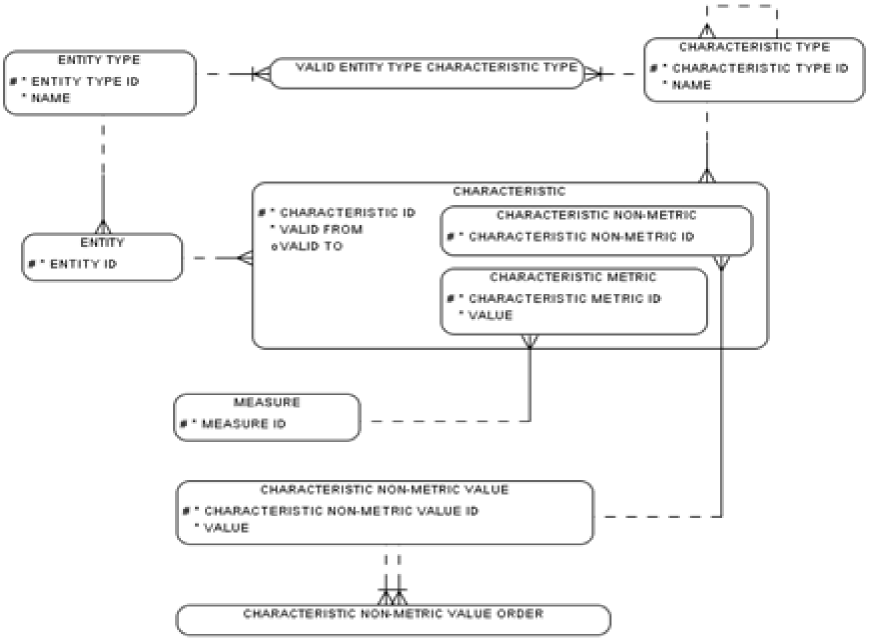

ENTITY TYPE
# * ENTITY TYPE ID
* NAME
VALID ENTITY TYPE CHARACTERISTIC TYPE
CHARACTERISTIC TYPE
# * CHARACTERISTIC TYPE ID
* NAME
ENTITY
# * ENTITY ID
CHARACTERISTIC
# * CHARACTERISTIC ID
* VALID FROM
o VALID TO
CHARACTERISTIC NON-METRIC
# * CHARACTERISTIC NON-METRIC ID
CHARACTERISTIC METRIC
# * CHARACTERISTIC METRIC ID
* VALUE
MEASURE
# * MEASURE ID
CHARACTERISTIC NON-METRIC VALUE
# * CHARACTERISTIC NON-METRIC VALUE ID
* VALUE
CHARACTERISTIC NON-METRIC VALUE ORDER

Figure 2. Highly abstracted characteristic pattern

.

Further on, different operational police applications will store different facts depending on their purposes and orientation. To be able to integrate all these operational crime data as well as to be flexible for integrating new criminological relevant characteristics we have developed the data model in Fig. 2.

The data model pattern in Fig. 2 is all about managing characteristics and can be applied to many different entities. It is based on the idea that a characteristic consists of attributes, e.g. the *color* of a car, and its specification, e.g. *red* [14].

In the pattern's center there is the CHARACTERISTIC representing a property. The CHARACTERISTIC is separated into METRIC and NON-METRIC depending on the kind of possible specifications. METRIC CHARACTERISTIC is used to manage properties with countable specification. In order to enable interpretations of distances, a METRIC CHARACTERISTIC is optionally related to a MEASURE. In contrast, NON-METRIC CHARACTERISTIC can be used to store uncountable measures. To avoid hidden redundancies, specifications are stored within a separate entity in this case.

The attributes VALID_FROM and VALID_TO take care of the time aspect of properties. Some properties may only be valid for a period of time, especially if they are roles. Other properties may be constant in property but changeable in specification, which certain timestamps can handle as well.

By using a power type like CHARACTERISTIC TYPE it is possible to add categories or hierarchies to

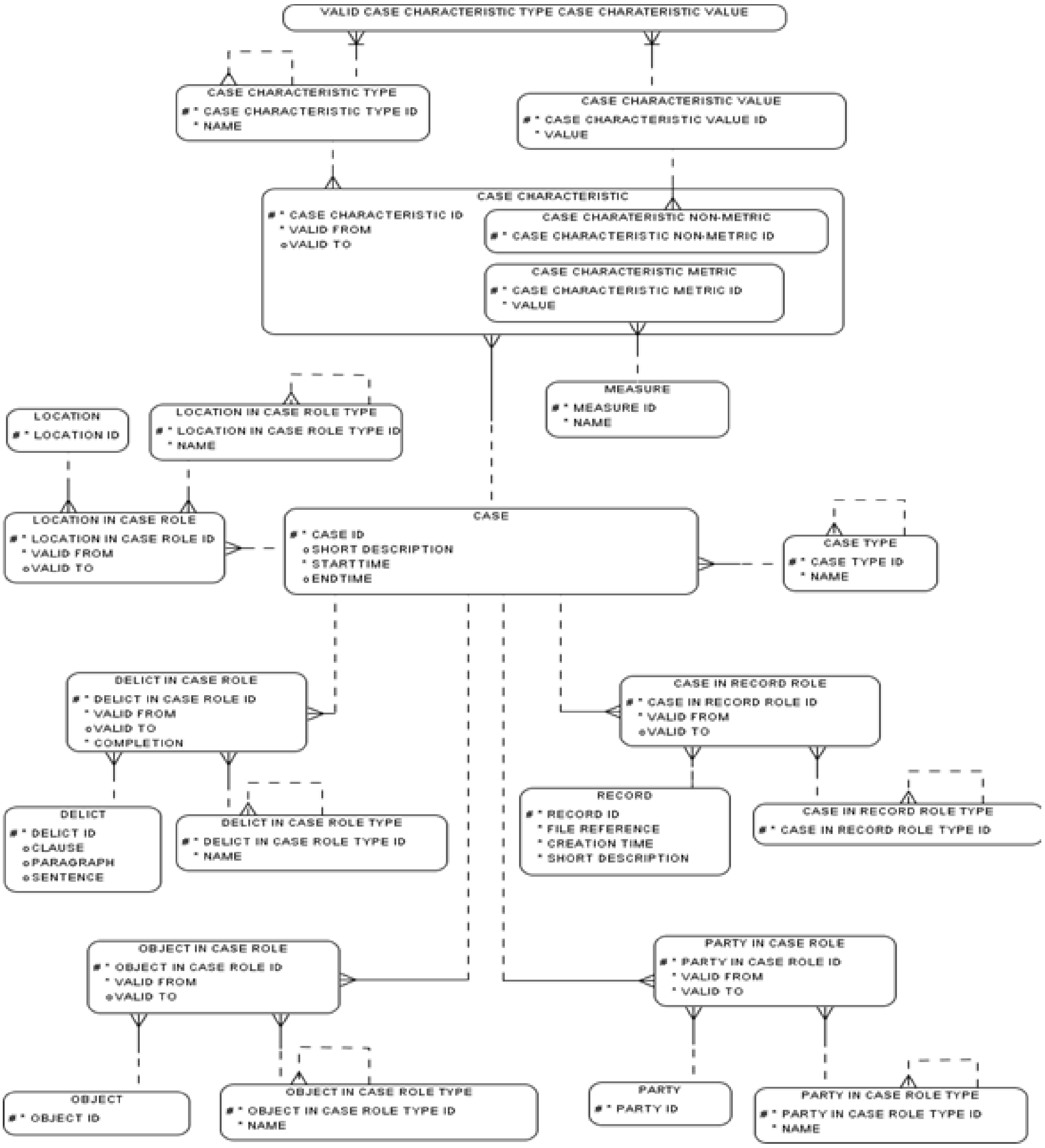


Figure 3. Cases and their contextual relationships

store characteristics, which is very useful especially for drilling-functions in data analysis.

By implementing this pattern, a data model can store affected entities in a much more flexible way. For example, decorating parties (this means persons, organizations, etc.) with characteristics enables you to store every single important fact about a suspect. This could be the body height, hair and eye color as well as clothing or behaviors. Every new characteristic can be easily applied by just adding a new characteristic type on data level without changing the models structure.

B. Records and Cases

Our data model about records and cases in Fig. 3 is less abstract in terms of different fields of application.

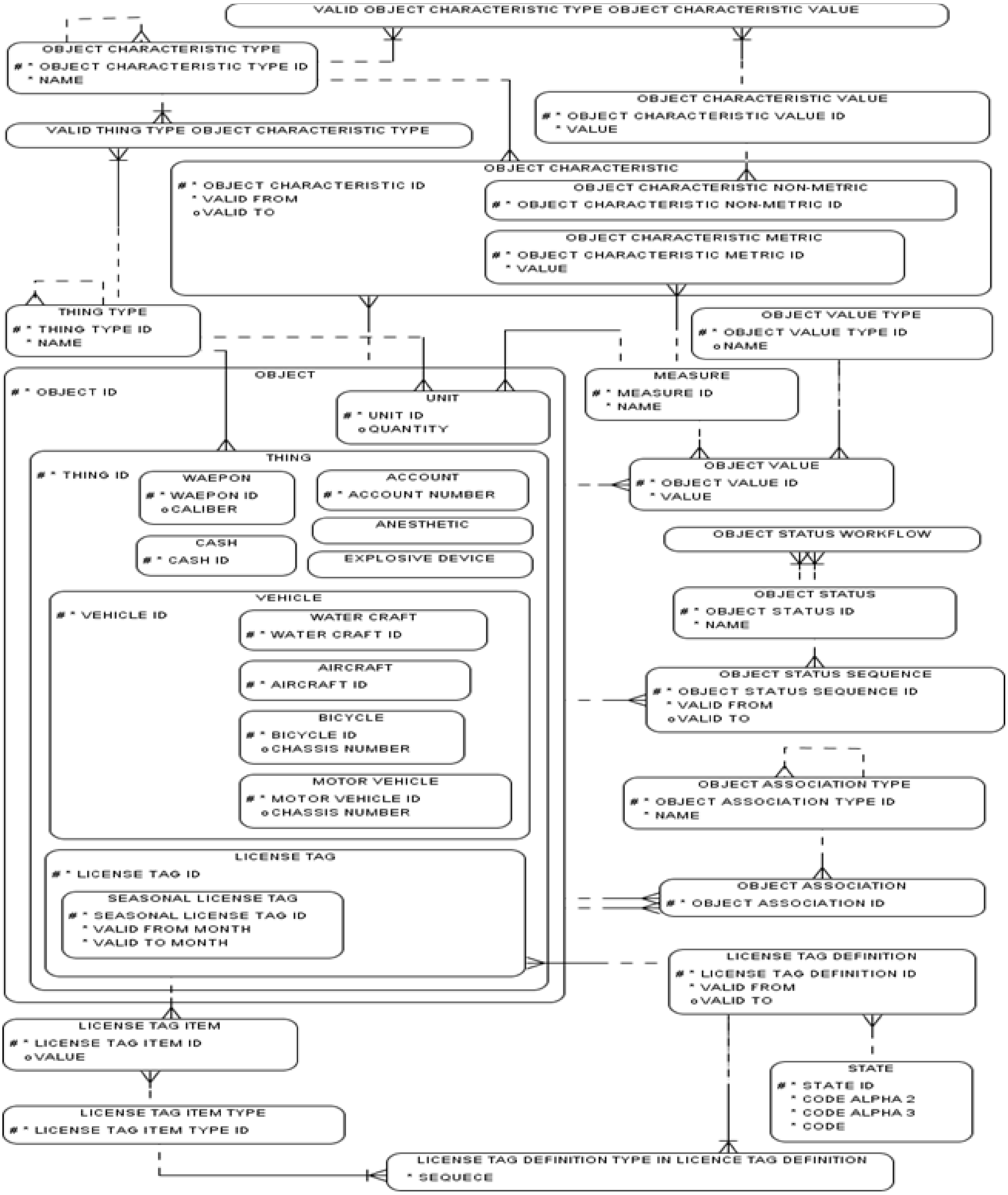


Figure 4. Things pattern

Nevertheless, it is built out of abstract data model patterns so that it is very flexible within the context of crime data.

A RECORD is the central element of operational police work. It is the representation of an old-fashioned file or document that bundles every relevant items or cases. It has no direct crime fighting significance but it is used to structure data on a higher level, especially in operational work.

More important for crime data analysis is the CASE entity instead. The CASE can be understood as every committed criminal action. Because of that, a CASE contains contextual relationships to nearly every other entity stored in the whole data model. For this reason, it has to be possible to store LOCATIONS as well as PARTIES or THINGS depending on their role.

This kind of modelling is what achieves the needed flexibility. By adding roles or entity types one can easily expand the model without structural changes.

We used the characteristic pattern (Fig. 2) within the case data model (Fig. 3) in order to store the way of committing crimes or other properties that might be unknown now.

Different from the characteristic model, this pattern mainly adds value by holding relationships instead of objects details. As we described in Section II, this is one of the key tasks for crime fighting data models. This data model allows storing a very detailed view of cases and their circumstances. Looking at a case one can extract relevant information like e.g. who was concerned, who is suspected of having committed an offence, or which things were used. This means the CASE entity enables allowing getting a full overview of a committed crime. From a technical perspective, this gets possible by using *contextual roles* which allows setting up one single relationship on a structural level that can be specified on the object level by using role types. Now we can store new kinds of relationships easily by just adding roles that concretize an existing relationship.

C. Things

Among the administrative level shown in Section IV B one of the main tasks in police work is managing things corresponding to a case. This may contain stolen or damaged goods, evidence or anything else. The main difference between the police and conventional business context is that criminological data processing has to be able to store nearly every kind of thing at a very high level of granularity. Thus, a data model managing things has to be very close to the real world. Remembering that one of the key steps in data model development is to define the relevant real world cutout, we now have to enable the database to store nearly any real world requirement.

Fig. 4 is an approach of meeting all the mentioned goals. In the center of this figure the key entity THING representing any possible physical item. To give a glue of what a thing can be, the entity exemplarily contains some derivatives like WEAPON, VEHICLE or ANESTHETIC. Of course, every sub-entity can be specialized further on if needed.

As the police not only deal with single items, the model has to enable the database to store UNITS. This is to be explained by means of an example: Now we can store one stolen notebook by using the entity THING as well as a whole bunch of stolen notebooks by using the entity UNIT. Because both entities are very similar in shaping and share many relationships to other entities, they are both generalized to OBJECT, which mainly is an expedient to hold relationships centrally.

In conventional modeling every sub-entity of THING would have to hold its nature properties. Still, this would be an acceptable and solid approach. Because of the variety, it would be hard to predefine these properties as well as to decide on the optimal abstraction level. This is why we decided to outsource properties of a thing to the already introduced pattern of characteristics.

The THING model is a very good example of how to make use of universal data model patterns like characteristics and the strength of generic data models. The model stays with conventional modeling and database techniques but is also very near to real world requirements. This also clarifies that multi-structured data is not necessarily needed in order to add flexibility to data models.

The THINGS pattern combines detailed object descriptions with many relationships to other entities and the fields of application are very wide spread.

V. CONCLUSION

The concepts of patterns have already been known for many years in software engineering. By adopting these ideas and transferring them to the process of data modeling especially regarding to the police domain, we can increase the quality of data models while simultaneously decreasing development costs. There is still a need of methodological background. The method of Generic Data Modeling enables the developer to integrate quality management into the process of modeling.

In this paper we illustrated the use of Generic Data Modeling in the context of data modeling for crime data. Generic Data Modelling allows us to develop flexible data models that scale with the agile requirements for crime fighting and offer the needed strength to be a solid platform for data analysis at the same time. Referring to the use of such a data model for police data warehouse, it becomes possible to integrate data from different heterogeneous sources easily. Because of the abstraction that is based on a target-oriented analysis of the problem domain instead of technical requirements this model can perfectly be used as a structural protection layer for different data marts.

Further research work needs to deal with demonstrating the performance of data analysis based on such abstract data models, quantifying the benefits of using patterns for data modeling and transferring Generic Data Modeling concepts to other application domains as well.

REFERENCES

[1] Inmon, W. H.: Building the data warehouse, 4th edition, Indianapolis (2005)

[2] Hartley, K., Seymour, L.F.: Towards a framework for the adoption of business intelligence in public sector organisations: the case of South Africa, in: Proceedings SA-ICSIT’11, pp. 116-122 (2011)

[3] Kim, G.-H., Trimi, S., Chung, J.-H.: Big data applications in the government sector, in: Communications of the ACM, Vol. 57, No. 3, pp. 78-85 (2014)

[4] Chen, H. et al.: Crime Data Mining: A General Framework and Some Examples, in: Computer, Vol. 37, No. 4, pp. 50-56 (2004)

[5] Ludewig, J., Lichter H.: Software Engineering, Dpunkt, Heidelberg (2010)

[6] Bauer, F. L.: Software Engineering. Wie alles begann, in: Informatik Spektrum, Nr. 5, pp. 259-260 (1993)

[7] Alexander, C.: A Pattern Language: Towns, Buildings, Constructions. Oxford University Press (1978)

[8] Gamma, E: Design patterns. Elements of reusable object oriented software. (1995)

[9] Hay, D. C.: Data model patterns. Conventions of thought, New York (1996)

[10] Silverston, L., Agnew, P.: The data model resource book, Indianapolis (2009)

[11] Schneider, S., Frosch-Wilke, D.: Analysis Pattern in Dimensional Data Modelling, in: Kannan, R., Andres, F. (ed.): Data Engineering and Management – Second International Conference, ICDEM 2010; LNCS Vol. 6411, Springer, pp. 109-116 (2012)

[12] Tešanovic, A.: What is a pattern?, URL: http://st.inf.tu-dresden.de/Lehre/dpf/IntroductoryPapers/tesanovic-WhatIsAPattern.pdf, April 24, 2014

[13] Schneider, S.: Konstruktion generischer Datenmodelle auf fachkonzeptioneller Ebene im betrieblichen Anwendungskontext. Methode und Studie. Dissertation. European Business School (2007)

[14] Summerford, J.: Neither Universals nor Nominalism. Kinds and the Problem of Universals, in Metaphyica, No. 5, pp. 101 – 126 (2003)

Finding Potential Threats in Several Security Targets for Eliciting Security Requirements

Haruhiko Kaiya
Kanagawa University
Hiratsuka, Japan
Email: kaiya@kanagawa-u.ac.jp

Shinpei Ogata
Shinshu University
Nagano, Japan
Email: ogata@cs.shinshu-u.ac.jp

Shinpei Hayashi
and Motoshi Saeki
Tokyo Institute of Technology
Tokyo, Japan
Email: {hayashi,saeki}@se.cs.titech.ac.jp

Takao Okubo
Institute of Information Securiry
(IISEC) Yokohama, Japan
okubo@iisec.ac.jp

Nobukazu Yoshioka
National Institue of Informatics
(NII) Tokyo, Japan
nobukazu@nii.ac.jp

Hironori Washizaki
Waseda University
Tokyo, Japan
washizaki@waseda.jp

Atsuo Hazeyama
Tokyo Gakugei University
Tokyo, Japan
hazeyama@u-gakugei.ac.jp

Abstract—**Threats to existing systems help requirements analysts to elicit security requirements for a new system similar to such systems because security requirements specify how to protect the system against threats and similar systems require similar means for protection. We propose a method of finding potential threats that can be used for eliciting security requirements for such a system. The method enables analysts to find additional security requirements when they have already elicited one or a few threats. The potential threats are derived from several security targets (STs) in the Common Criteria. An ST contains knowledge related to security requirements such as threats and objectives. It also contains their explicit relationships. In addition, individual objectives are explicitly related to the set of means for protection, which are commonly used in any STs. Because we focus on such means to find potential threats, our method can be applied to STs written in any languages, such as English or French. We applied and evaluated our method to three different domains. In our evaluation, we enumerated all threat pairs in each domain. We then predicted whether a threat and another in each pair respectively threaten the same requirement according to the method. The recall of the prediction was more than 70% and the precision was 20 to 40% in three domains.**

Keywords–Security Requirements Analysis; Requirements Elicitation; Common Criteria; Security Target; Domain Knowledge.

I. Introduction

Knowledge about computer security is important for requirements elicitation because security has effects on development costs and efforts. Although we expect that security experts will provide such knowledge, they cannot always do so. Researchers thus developed methods of eliciting requirements using documented knowledge [1] [2] [3] [4]. In such methods, the method helps a requirements analyst to find new security requirements on the basis of requirements already elicited. However, there are a few methods of developing or acquiring such knowledge [5] [6].

It was not easy to acquire such security knowledge that are high quality because security experts tacitly held the knowledge and they rarely document it. There have recently been several structured documents that have been of high quality where such knowledge has explicitly been represented. Examples are the Security Target (ST) in Common Criteria (CC) [7] and Common Attack Patterns Enumeration and Classification (CAPEC) [8]. Each element in a document is uniquely identified in such documents, and the relationships between the elements are formally specified. Existing methods of developing documented knowledge did not fully utilise the explicit structure of knowledge sources, but they simply used linguistic characteristics in such documents with the help of lightweight natural language processing (NLP).

Saeki et al. [9] reported that using more than two knowledge sources contributed to comprehensively eliciting security requirements. Especially, when a threat in a ST and similar threat in another ST were together examined in our current requirements analyses, security requirements could be elicited more comprehensively than ever. However, it took a huge amount of effort because useful knowledge was scattered over several different sources, and requirements analysts had to manually find them step by step. Therefore, such sources have to be integrated so that the analysts can efficiently and comprehensively find potential threats protected by security requirements. However, no one cannot know a threat and another will threaten the same requirement without examining their contents.

We then set up three research questions.

- RQ1: How to integrate several structured security documents systematically so that requirements analysts can elicit security requirements comprehensively and efficiently?
- RQ2: How to use security documents written in different languages, such as English and French?
- RQ3: How to perform such integration without the knowledge whether one threat and another will respectively threaten the same requirement?

The contribution of this paper is to provide answers to research questions above.

We propose a method of integrating several STs in this paper for three main reasons. First, ST provides highly structured documents that are useful for semantically integrating them. Second, the documents are provided in machine-readable format. Third, STs refer to issues related to requirements although most security related documents refer to design and/or implementation issues.

We assume that analysts first find partial threats to a system by themselves, and the integrated STs contain them. Countermeasures to the threats are candidates for the security requirements in the system in our method. Our method then recommends the other threats to be examined for eliciting additional security requirements. It is very important to find potential candidates for threats when security requirements are being elicited because threats that are not taken into account make the system vulnerable. The integrated STs thus contribute to improving the quality of eliciting security requirements because they provide as many candidates as possible.

The relationships between a threat and other threats should be systematically identified to find such additional threats. Various kinds of semantic relationships such as similarities and dependencies between threats are useful. We focus on the means for protection against each threat to identify these relationships. Such means are called security functional requirements (SFRs) in CC. We assume several SFRs are commonly used if such semantic relationships between a threat and another are established. Our method is based on this assumption. The assumption is confirmed and explained in Section V.

The rest of this paper is organised as follows. The next section reviews related work on methods of eliciting requirements using reusable knowledge, and methods of developing such knowledge. Section III introduces CC, which provide structured knowledge on security, and methods using that knowledge. Section IV presents a method of integrating several STs. We also present its background, requirements and a discussion. We next explain our evaluation on whether our method worked well in Section V. We address the three research questions in summary and pose future issues.

II. Related Work

Structured knowledge, in general, such as ontology is widely used in the field of software engineering [10]. There are numerous methods of eliciting requirements using ontology in the field of requirements engineering [1] [2] [3] [4]. We expect knowledge such as ontology will contribute to the completeness and correctness of requirements, and such an expectation will only be satisfied when this knowledge is comprehensive enough. However, it is not easy to develop or acquire such knowledge in real situations. We thus have to investigate how to develop or acquire such knowledge.

Two types of different researches related to ontology exist. One is about designing the meta-model (syntax) of the ontology [11] [12]. Another is about developing the concrete instances of ontology. Our method in this paper belongs to the second type, and we simply use the structure of the ST as a meta-model of model instances.

Research on developing ontology in general already exists [13] [14] [15] [16], and most researchers have used NLP techniques. There are already a few knowledge integration methods and tools for requirements engineering [5] [6], and most of them also use NLP, and do not focus on highly structured documents. A method of developing security ontology was also proposed [17]. However, few automated or reused mechanisms were taken into account in the method.

Much structured knowledge is available in the field of security. However, most of it is not related to requirements but to design or implementation [18]. The ST in CC explained in detail below is one of a few exceptions because requirements related concepts such as threats and objectives are highlighted in it.

III. CC and Methods of Eliciting Requirements Using their Knowledge

CC represents an international standard that prescribes how to write documents to assess the security properties of information systems. CC consists of several structured documents and one of these is the ST. ST can be used to improve security requirements elicitation for the following reasons. First, we can find explicit relationships among threats to assets in a system, objectives to mitigate or avoid individual threats, and functionalities to implement the objectives, which are called SFRs. We can usually find the list of assets and implicit relationships between the assets and other elements above. Second, we can find already certified STs for individual IT products [7] and they are categorised on the basis of the types of the products.

Figure 1 outlines part of an ST for an Information Technology (IT) product for an Integrated Circuit (IC) card, where each threat and objective has its own unique name such as T.Skimming and O.Data_Conf, and each of them has its own explanation written in natural language sentences. The templates of SFRs are chosen from the catalogue provided by the CC framework (called CC Part 2) with several parameters, and each template is instantiated in each ST by assigning some values to the parameters. Each SFR has its own unique name such as FIA_UID or FCS_COP. The explicit relationships among threats, objectives and SFRs are provided in the ST so that each threat is countered by some concrete means, i.e., SFRs. The relationships between assets and the others are implicitly provided because no formal rules for naming assets and relating them to others are specified in CC. The reader of an ST should manually follow and understand issues related to assets.

Because the knowledge in STs is useful and it is easy for computers to operate them, as previously mentioned, there are already several methods of requirements elicitation [9] [19] [20] [21] using STs. In most of them, elements such as threats, objectives and SFRs are used for resources to elicit security requirements on the basis of existing functional requirements or goals.

We briefly introduce a method of eliciting security requirements using the knowledge in STs in a previous article [9]. Figure 2 outlines the flows of inputs and outputs with the method. The inputs are knowledge and functional requirements (FRs), and the outputs are security requirements (SRs). We can explain the steps in the method with this figure. We can also explain why integrating several STs is helpful in the method.

1) An analyst elicits or acquires functional requirements (FRs) in advance. There is “FR10” at the top right of Figure 2.
2) He/she has to explore assets in an FR or its threats by referring to descriptions of assets and threats in an ST. In this example, he focuses on the term “IC card” in the FR because an ST C229 contains an asset called an “IC chip”.

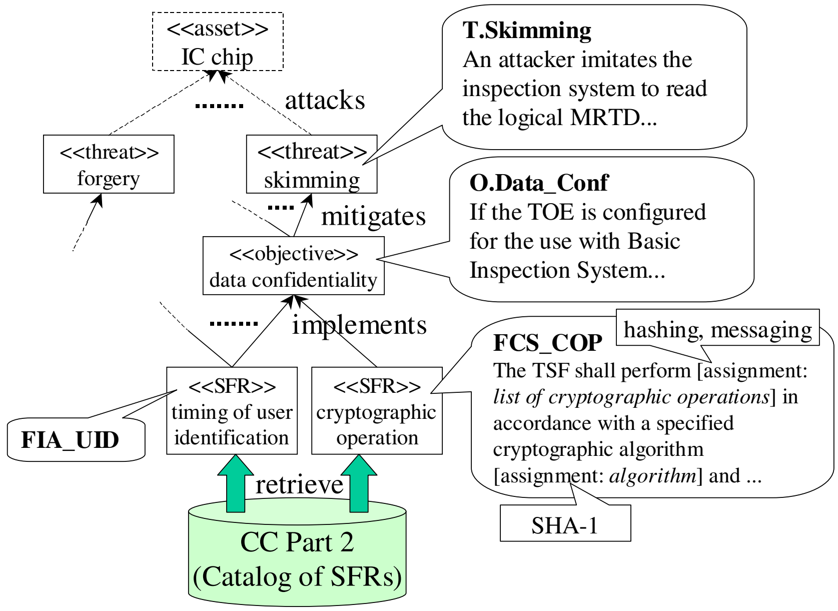

<<asset>>
IC chip
attacks
<<threat>>
forgery
<<threat>>
skimming
T.Skimming
An attacker imitates the inspection system to read the logical MRTD...
mitigates
<<objective>>
data confidentiality
O.Data_Conf
If the TOE is configured for the use with Basic Inspection System...
implements
<<SFR>>
timing of user identification
<<SFR>>
cryptographic operation
FIA_UID
FCS_COP
hashing, messaging
The TSF shall perform [assignment: list of cryptographic operations] in accordance with a specified cryptographic algorithm [assignment: algorithm] and ...
SHA-1
retrieve
CC Part 2
(Catalog of SFRs)

Figure 1. A part of ST for IC Card System

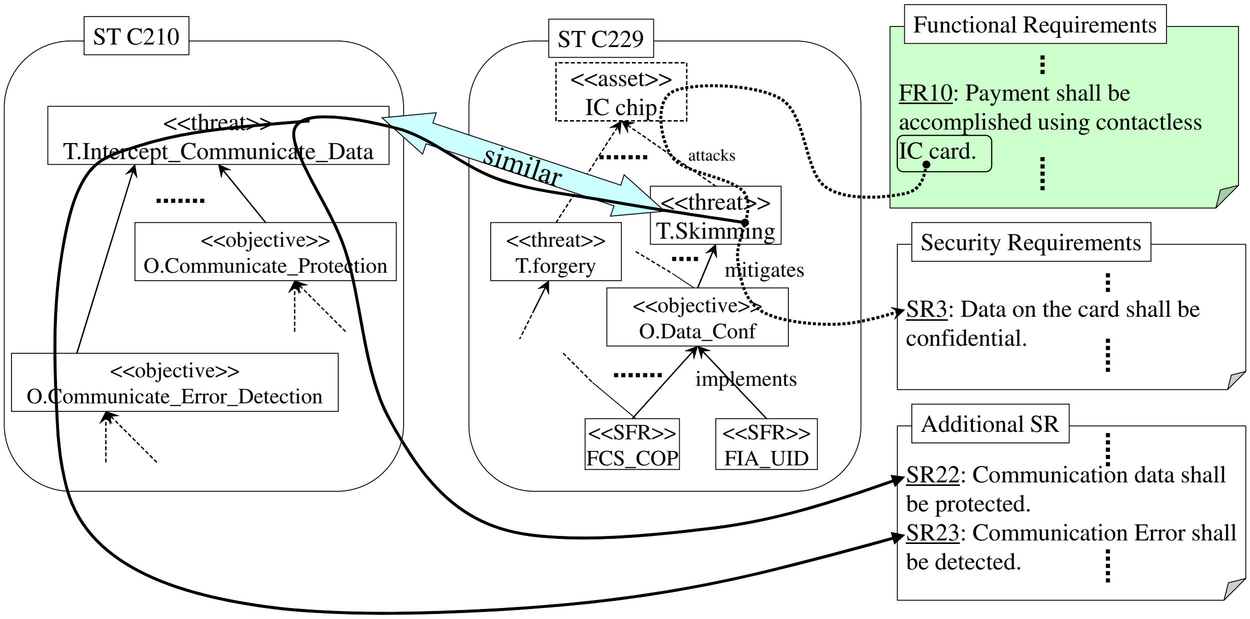


Figure 2. Example of security requirements elicitation using two STs

3) He/she then identifies threats to the asset, and regards objectives to the threats in the ST as candidates of security requirements (SRs).

   In this example, an objective "O.Data_Conf" is added as "SR3" because the objective mitigates the threat "T.Skimming" and the threat threatens the asset "IC chip" in the ST. The dotted curved line in the figure indicates the trace in this step.

4) He/she finds threats in other STs that are similar to the threats that were originally identified. He/she then regards objectives to the threats in other STs as additional candidates of SRs.

   He/she assumed "T.Skimming" in C229 in Figure 2 was similar to "T.Intercept_Communicatie_Data" in C210. He/she thus systematically finds SR22 and SR23 as seen in the figure. The curved lines in the figure indicate the traces in this step.

5) He/she repeats the steps above for each FR.

Step 4 plays a role in integrating several different STs, but how to integrate them is beyond the scope of Saeki et al. [9]. The main goal of the research discussed in this paper is to identify similarity among threats used in Step 4.

IV. Similar threats derived from several STs

The goal of the method presented in this paper is to derive the pair of similar threats each of which threatens the same requirement from several STs. A pair of "T.Intercept_Communicate_Data" and "T.Skimming" in Figure 2 is an example of this pair. We predict such a pair on the basis of SFRs commonly mitigating or avoiding both threats. Even if several SFRs are common countermeasures to two threats, the threats are not always similar to each other. We thus need *a threshold* on such commonality so that we determine whether two threats with a certain commonality are similar or not. The threshold should be determined without the knowledge whether two threats are actually similar to each other. Metrics in Section IV-B are mainly used to explain how to determine this threshold, and how to derive pairs of threats is explained in Section IV-C.

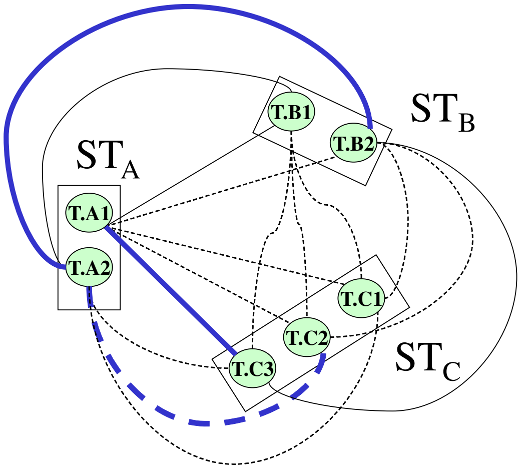
A. Background and Requirements for Deriving Threat Pairs

The method of eliciting security requirements in Section III uses relationships among existing FRs, attackers requirements (threats), countermeasures (objectives) and concrete means for implementing countermeasures (SFRs). Although relationships among threats, objectives and SFRs are explicitly represented in STs, requirements analysts should identify the relationships between FRs and others by themselves. Focusing on assets in individual FRs is one way to make such identifications, but we have to take into account synonyms are handled for the name of an asset. In addition, relationships between assets and threats should be manually investigated in an ST as was mentioned in Section III. Even though there are difficulties in identifying the relationships between FRs and others, an analyst has to first find one or a few threats, objectives or SFRs related to an FR.

Once one or a few of them can be found, the explicit structure in an ST like that in Figure 2 can be systematically utilized to expand the number of candidates of threats to each FR. Finding STs in the same application domain is not very difficult because they are categorised into domains on their web site [7]. Establishing a relationship between a threat in an ST and another threat in another ST is not very easy if we simply read their explanatory sentences. Although terms in an ST are normally used consistently within an ST, there is little consistency in terms between one ST and another, especially individual STs have been written by different technical people or different companies. In addition, STs can be written in different languages such as English, Japanese or French. In such cases, it is very difficult to integrate several different STs on the basis of terms in each ST.

B. Metrics related to Finding Threat Pairs

We define the following metrics for the method of deriving pairs of similar threats below. Let us use the example in Figure 3 to explain each metric. Each box in this figure corresponds to an ST, each circle corresponds to a threat, and each straight or dashed line specifies a potential pair explained below. The number of all potential pairs is 16. A legend for each kind of line can be found in the figure. In this example, threat pairs are derived from three STs. The STs are called $\mathrm{ST_A}, \mathrm{ST_B}, \mathrm{ST_C}$, where $\mathrm{ST_A}$ contains two threats, $\mathrm{ST_B}$ contains two threats, and $\mathrm{ST_C}$ contains three threats respectively.



Legend & summary

Correct \ Estimated	Yes	No
Yes	2 ━━━	1 ━ ━ ━
No	3 ───	10 ┄┄┄

Figure 3. Example of three STs and threat pairs among them.

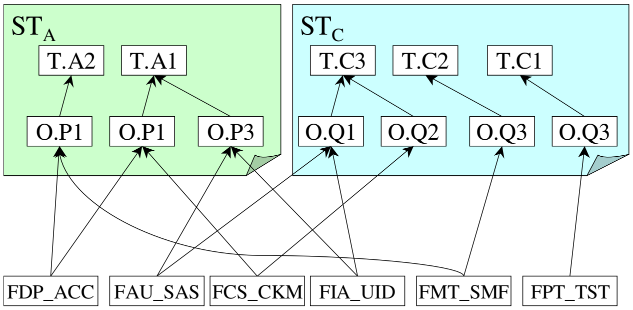


Figure 4. Example of two STs to explain metrics in Section IV-B

- *Potential pairs*:
  Let us focus on a pair of threats, each of which belongs to different STs. We call all such pairs *potential pairs*, and POT denotes the set of all potential pairs. There are 16 potential pairs, $|POT|$, in Figure 3. The number is derived on the basis of (1).

$$(\sum_{\mathrm{i=A}}^{\mathrm{C}}(\mathrm{nt}(\mathrm{ST_i}) \times ((\sum_{\mathrm{j=A}}^{\mathrm{C}}\mathrm{nt}(\mathrm{ST_j})) - \mathrm{nt}(\mathrm{ST_i}))))/2 \quad (1)$$

  In (1), $\mathrm{nt}(x)$ is a function to obtain the number of threats in x, which is the name of an ST. The equation (1) is instantiated in Figure 3 as follows because $\sum_{\mathrm{j=A}}^{\mathrm{C}} \mathrm{nt}(\mathrm{ST_j})$ is seven.

$$((2 \times (7-2)) + (2 \times (7-2)) + (3 \times (7-3)))/2$$

- *Correct pairs*:
  As was explained in the introduction, we expect threat pairs will help a requirements analyst to elicit additional security requirements when he/she finds one or a few threats threaten existing requirements. If both threats in a pair respectively threaten the same

requirement, we call the pair *a correct pair*, and the set of all correct pairs is represented as COR. The correct pairs are depicted by the thick line in Figure 3. Any correct pair is contained in the potential pairs, i.e., $COR \subseteq POT$. We cannot know COR in an actual situation because the number of POT is usually too huge to examine the meaning of each threat.

- *Estimated pairs*:
  Because no one knows the correct pairs, we have to predict whether a pair is a correct pair or not on the basis of computer-identifiable information. We used *SFR-commonality* in our method, which is explained below, to make this prediction. When SFR-comonality determines that a potential pair seems to be a correct pair, we call the potential pair *an estimated pair*, and the set of all estimated pairs is called EST. Any estimated pair is contained in potential pairs, i.e., $EST \subseteq POT$. Estimated pairs are represented by straight lines in Figure 3. The decision by SFR-commonality is not always correct. For example, there are five estimated pairs and only two out of them are correct in Figure 3. In total, 12 out of 16 pairs are correctly decided, but the others are not.
- *SFR-commonality*:
  This commonality is used to determine whether a threat in a ST is similar to another in another ST. It is a function with a threat pair as an argument. When the value is nearby one, two threats are similar to each other. When the value is nearby zero, they are not similar. We use the Jaccard index [22] to define SFR-commonality. The Jaccard index Jac returns the degree of commonality between two sets X and Y as shown in (2).

  $$Jac(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{2}$$

  The returned degree of (2) takes a value from zero to one. To use the Jaccard index to construct SFR-commonality, we define the following function (3).

  $$USFR : T \rightarrow 2^{ALLSFR} \tag{3}$$

  T in (3) is the set of threats and $ALLSFR$ is the set of all existing SFRs defined in CC Part II. Function $USFR$ returns the set of SFRs transitively used in a threat t. We will now explain how to calculate the commonality between two threats by using the example in Figure 4, in which two STs, threats, objectives and SFRs used in the STs are depicted. Intuitively, T.A1 and T.C3 are similar to each other because almost the same SFRs were used to avoid or mitigate threats. Commonality is defined here so that it represents such intuition.
  The results obtained by applying $USFR$ to the threats in Figure 4 are:
  $USFR(\text{T.A2}) = \{\text{FDP\_ACC}, \text{FMT\_SMF}\}$
  $USFR(\text{T.A1})$
  $= \{\text{FDP\_ACC}, \text{FAU\_SAS}, \text{FCS\_CKM}, \text{FIA\_UID}\}$
  $USFR(\text{T.C3})$
  $= \{\text{FAU\_SAS}, \text{FCS\_CKM}, \text{FIA\_UID}\}$
  $USFR(\text{T.C2}) = \{\text{FMT\_SMF}\}$
  $USFR(\text{T.C1}) = \{\text{FPT\_TST}\}$
  Because each potential pair of threats p consists of two threats $t1$ and $t2$, SFR-commonality Com can be defined as shown in (4).

  $$Com(p) = Jac(USFR(t1), USFR(t2)) \tag{4}$$

  In (4), p is $\{t1, t2\}$.
  Because whether a pair is an estimated pair depends on the threshold td mentioned above, the set of estimated pairs EST is parameterised by the threshold, i.e., $EST(td)$, and is defined as shown in (5).

  $$EST(td) = \{x | x \in POT \wedge Com(x) \geq td\} \tag{5}$$

  In (5), POT is the set of all potential pairs.
  For example, we calculate SFR-commonality in Figure 4 as:
  $Com(\{\text{T.A1}, \text{T.C1}\}) = 0$
  $Com(\{\text{T.A1}, \text{T.C2}\}) = 0$
  $Com(\{\text{T.A1}, \text{T.C3}\}) = 3/4 = 0.75$
  $Com(\{\text{T.A2}, \text{T.C1}\}) = 0$
  $Com(\{\text{T.A2}, \text{T.C2}\}) = 1/2 = 0.5$
  $Com(\{\text{T.A2}, \text{T.C3}\}) = 0$
  We can then calculate $EST()$ as:
  $EST(0.1)$ = { {T.A2, T.C2}, {T.A1, T.C3} }
  $EST(0.5)$ = { {T.A2, T.C2}, {T.A1, T.C3} }
  $EST(0.7)$ = { {T.A1, T.C3} }
  $EST(0.9)$ = { }
- *Estimated gain*:
  We expect a threat pair will suggest an additional threat protected by additional security requirement(s) when a threat in the pair has already been identified. Therefore, the number of the threats used for finding security requirements has increased by the number of threat pairs. The estimated gain indicates such a degree, and is defined as in (6).

  $$\frac{|EST| + |ALL|}{|ALL|} \tag{6}$$

  In (6), ALL is the set of all threats. In Figure 3, ALL is {T.A1, T.A2, T.B1, T.B2, T.C1, T.C2, T.C3}, and the estimated gain is about 1.7 (=((2+3)+(2+2+3))/(2+2+3)). We occasionally represent this rate as a percentage such as 170 %. When the estimated pairs are rigorously chosen, $|EST|$ becomes nearly zero. In that case, the estimated gain becomes nearly 100 %. We do not apply transitivity when we calculate gain. For example, the pair T.A1 and T.B2 in Figure 3 is a transitive estimated pair because the pair T.A1 and T.C3 is an estimated pair and the pair T.C3 and T.B2 is also. However, we do not apply transitivity because most threats can become pairs with such transitivity.
  Because we can calculate the estimated gain in actual cases, we expect that the estimated gain can be an indicator to define the SFR-commonality below. The reasons for the expectation are as follows. Because of $EST \subseteq POT$, the upper limit of the estimated gain can be known in advance, i.e., $\frac{|POT|+|ALL|}{|ALL|}$. Because $EST = POT$ is unrealistic, the estimated gain should not become this upper limit. When an ST contains a set of threats, another similar ST normally contains a similar set. The total number of threats, i.e., $|ALL|$,

thus influences the proper size of EST, i.e., the size of COR, because of the commonality of sets of threats between STs. We thus expect that the estimated gain will be an indicator to predict estimated pairs.

C. Method of deriving Threat Pairs

The goal of our method is to derive threat pairs from several STs so that requirements analysts can elicit security requirements against the potential threats. The main characteristics of this method are explained in what follows.

- The method enables us to use several STs written in different languages such as English and French together because STs use common SFRs.
- The method can be automatically executed except to determine the threshold. The threshold is used to determine whether one threat can be regarded to be similar to another.
- In the method, the change of estimated gain plays a role of a clue to determine the threshold. The change of estimated gain can be automatically calculated without knowledge as to whether two threats are actually similar to each other.

The six steps in the method are as follows. All steps except step 5 can be performed automatically.

1) We prepare several STs that belong to the same domain. Although only three STs were integrated in our evaluation, our method can accept any number of STs and there are a huge number of STs in the same domain. For example, we could find 653 STs in the IC card domain [7].
2) We enumerate POT in these STs.
3) We calculate the SFR-commonality for each potential pair. It takes a value from zero to one.
4) We have to define a threshold on SFR-commonality so that we determine EST. We call a candidate for such a threshold *the lower limit of commonality*. We present the changes in estimated gain along with progress in the lower limit of commonality.
5) We choose a lower limit as *the threshold*. The threshold should be subjectively determined, but the changes in the estimated gain along with the progress in the lower limit will help us to determine this.
   A typical way to determine the threshold is as follows. An analyst focuses on the estimated gain where the lower limit is zero. The estimated gain in this case has the largest value, i.e., $(|POT| + |ALL|)/|ALL|$. He/she then increases the lower limit step by step until the estimated gain becomes stable. The lower limit at the beginning of the stable range can be a threshold.
6) EST is determined by the threshold above. According to each threat pair in EST, a threat in the pair is recommended as resources to expand security requirements when a requirements analyst has already elicited one or a few security requirements protecting another threat in the pair.

D. Discussion

As was discussed in Section IV-A, we want to expand the candidates for threats to individual FRs. However, finding the first candidate is beyond the scope of this method.

We assume SFR-commonality is more helpful than the similarity based on the co-occurrences of words/terms in threats because of the problems mentioned in Section IV-A. Each threat in ST is mitigated/avoided by the set of security functions, i.e., SFRs. SFR-commonality focuses on the co-existences of means for such mitigation. There are generally several alternative means to satisfy a requirement, and a threat, which is a kind of attacker's requirement, is also satisfied in the same way. When attackers want to threaten assets, they at least utilise parts of similar or the same vulnerabilities, which correspond to weaknesses of assets [23]. Because the limited means against such vulnerabilities, i.e., SFRs, are known at least in CC, using SFRs to identify similarities in our method seems to be a good rationale. This issue is empirically confirmed and explained in the next section.

We compare the Jaccard, Dice and Simpson indices to calculate appropriate commonality between two sets. The Dice and Simpson indices are defined in (7) and (8).

$$Dice = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{7}$$

$$Simpson = \frac{|X \cap Y|}{min(|X|, |Y|)} \tag{8}$$

The Dice index takes almost the same value as the Jaccard index. The Simpson index is not suitable for pairs of sets, where the sizes of sets vary greatly. We thus used the Jaccard index in our method.

V. Evaluation

A. Preparation

We tested and confirmed that the method in Section IV-C worked well. We expect the method contributes to deriving threat pairs useful for expanding the candidates of security requirements. We thus focus on the recall and precision of EST. Because EST depends on a threshold for SFR-commonality and we subjectively determine the threshold by examining estimated gain, we especially confirm that estimated gain is a good clue to determine the threshold.

We use the following metrics in addition to the metrics in Section IV-B for our evaluation.

- *Recall*:
  Recall here refers to the degree to which how the decision by SFR-commonality can find as many correct pairs as possible. Recall can be defined in (9).

  $$\frac{|COR \cap EST|}{|COR|} \tag{9}$$

  In (9), $|S|$ is the number of elements in a set, S. The recall is about 0.66 (=2/3) in Figure 3. We expect recall will become 1.0 as much as possible because we want to find as many candidates for security requirements as possible.
- *Precision*:
  Precision here refers to the degree to which the decision by SFR-commonality is precise. Precision can be defined in (10).

  $$\frac{|COR \cap EST|}{|EST|} \tag{10}$$

TABLE I. STS IN OS DOMAIN

ID	No.# of Threats	Summary
MacOS	3	Apple Mac OS 10.6, Dec. 2009
RedHat	7	Red Hat Enterprise Linux Ver. 5.3 for CAPP Compliance on Dell 11th Generation PowerEdge Servers, Dec. 2009
Linux	13	Wind River Linux Secure 1.0, Apr. 2011

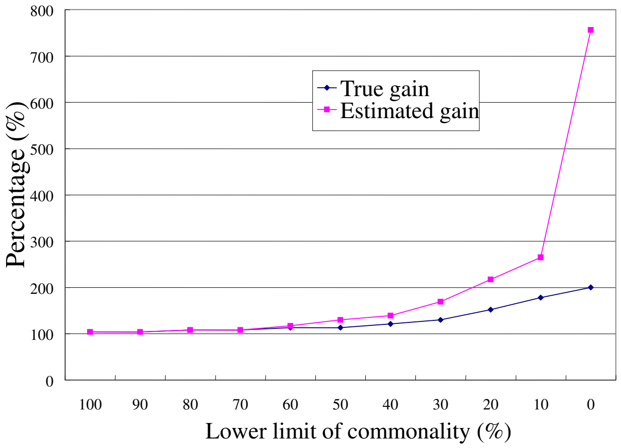


Figure 5. Changes in true and estimated gain of threat pairs in OS domain.

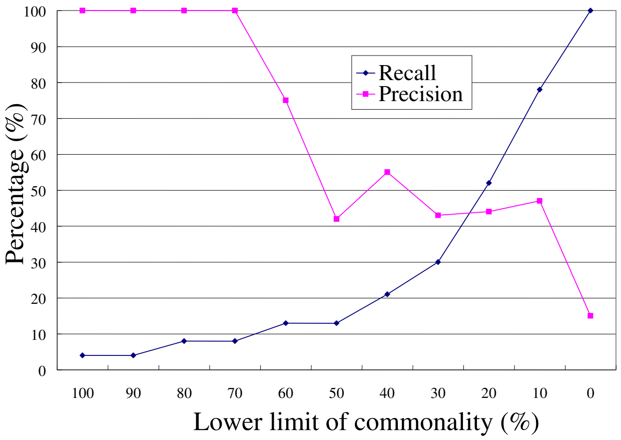


Figure 6. Changes in recall and precision of threat pairs in OS domain.

The precision in Figure 3 is 0.4 (=2/5). We also expect precision will become 1.0 as much as possible because we do not want to investigate threats that are not necessary. However, we can accept a certain degree of imprecision because precision and recall have trade-offs.

- *True gain*:
  Apparently, not all estimated pairs give us useful suggestion for finding additional candidates for security requirements because not all are correct pairs. We thus want to know the actual gain caused by estimated pairs if we can identify the correct pairs. The true gain indicates such a degree, and is defined in (11).

$$\frac{|EST \cap COR| + |ALL|}{|ALL|} \tag{11}$$

  The true gain in Figure 3 is about 1.3 (=(2+7)/7). We, of course, expect the true gain to be sufficiently large because missing threats to be investigated are avoided.

Because COR is necessary for evaluation, one author and his student decided it. They checked all potential threat pairs whether a threat and another in each pair respectively threatened the same type of functional requirements.

B. *Results*

We observed changes in the gains in operating systems (OSs) according to the changes in the lower limit of SFR commonality in Figure 5. The STs are summarised in Table I. Because there were a total of 23 threats and there were 23 correct pairs, the upper limit of true gain is 200 % as shown in the figure. Because there were a total 23 threats and there were 151 potential pairs, the upper limit of estimated gain was about 750 % (100*(23+151)/23) as well. As we can see from Figure 5, estimated gain largely increased when the limit changed from 10 to 0 %. According to our method presented in Section IV-C, 10 % was a suitable lower limit for SFR-commonality, i.e., the threshold. Because the precision and recall in Figure 6 were about 45% and 80% at the 10 % lower limit, estimated gain seemed to be a good clue to determine a suitable threshold.

We next used three STs in IC chip domain. Three are summarized in Table II. We used STs written in either English or Japanese in this domain. Because SFR-commonality is independent of the language used in individual STs, we could use such STs together. There were number 174 potential pairs, and 15 correct pairs. We plotted the changes in true and estimated gain of threat pairs in Figure 7 in the same way as that in Figure 5. The most suitable lower limit also seemed to be 10 % in this domain. We then plotted the changes in recall and precision in Figure 8 in the same way as that in Figure 6. Recall became sufficiently accurate (80%) at this lower limit according to this figure, while precision was acceptable (about 22%). The results indicated that most correct pairs could be estimated, and the errors in the estimates were acceptable. We also regarded our method as working well in this IC domain.

We finally used three STs in the domain of multi-function devices (MFDs). MFDs are also called multi-function printers (MFPs). The three STs are summarised in Table III. We used STs written in Japanese in this domain. There were 24 potential pairs, and four correct pairs. According to the changes in estimated gain in Figure 9, the most suitable lower limit was 20 or 10 %. As we can see from in Figure 10, recall had been about 75% from the 100 to 10 % lower limit. One of the reasons for this was that there were not that many correct pairs in this domain. This was one of the reasons recall has been good. Precision decreased from the 10 % lower limit, as indicated in Figure 10. These results in Figure 10 indicated that 20 % lower limit was the best one. This could

TABLE II. STS IN IC CARD DOMAIN.

ID	No.# of Threats	Summary
C191	6	IC card for residents, Oct. 2008. in Japanese
C210	9	Farmware for Mobile FeliCa IC chip, Feb. 2009, in Japanese
C229	8	Apollo OS e-Passport V1.0, Jul. 2009, in English

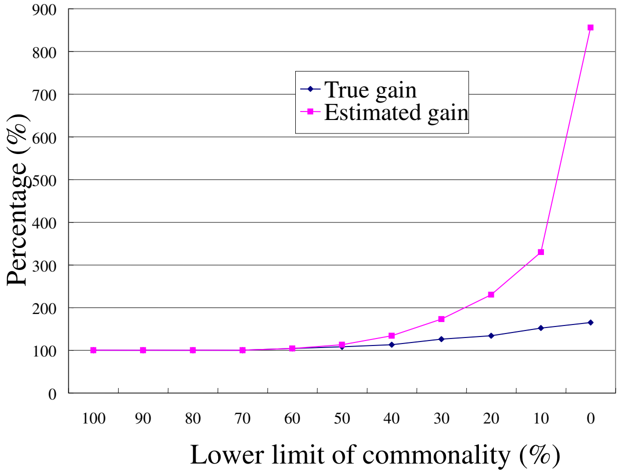


Figure 7. Changes in true and estimated gain of threat pairs in IC domain.

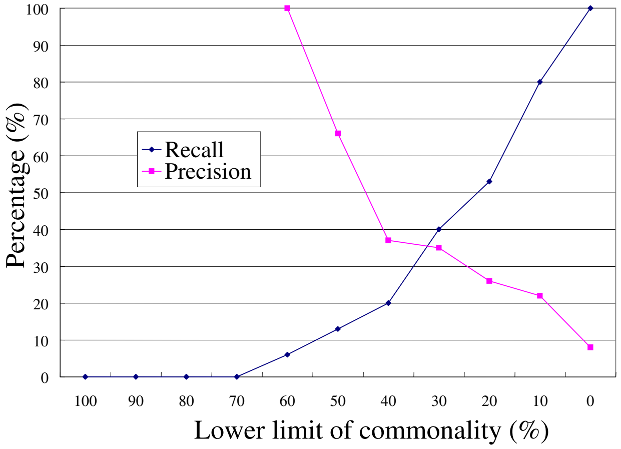


Figure 8. Changes in recall and precision of threat pairs in IC domain.

be estimated from the change in estimated gain in Figure 9 explained above. We also regarded our method worked well in this MFD domain.

C. Discussion

The method seemed to work well because the threshold could be decided by observing the changes of estimated gain. For each domain in this evaluation, the estimated gain suddenly and largely increased when the lower limit of commonality was nearby zero. We regarded the lower limit of commonality just before largely increasing as the threshold. At the threshold, the recall was more than 70% for each domain while the precision was 20 to 40 %. Because the goal of this method is to find potential threats as much and efficiently as possible, we may regard the results were acceptable.

We will now discuss the threats to validity in the evaluation. Most metrics in the evaluation depended on both estimated and correct pairs. Because the estimated pairs and correct pairs were defined separately, we did not worry about threats to internal validity. We used three different application domains for our evaluation, and the STs in each domain were written in two different languages, only in English, in English and Japanese, and only in Japanese. We thus regarded we had taken care of the problems caused by external validity. When the total number of STs is not very large, a requirements analyst does not have to use our method but simply investigates all the threats. Because it takes considerable effort to find correct pairs, only three STs were integrated in each domain and only 10 to 20 threats were included in the evaluation. There is no upper limit of the number on STs when using our method in actual situations because we do not have to find correct pairs and we can systematically calculate estimated pairs and gain as was explained in Section IV-C. We can thus derive a huge number of threat pairs where no analyst can investigate all the threats. The metrics in Section IV-B and Section V-A seem to be reasonable for measuring what we want to know, and most of them are systematically derived from the ST documents. However, only correct pairs should be subjectively determined. The correct pairs in our evaluation are determined at least two researchers so that we decreased threats to construct validity. Transitivity of pairs is not applied when true and estimated gains were calculated, which was also explained in Section IV-B. Therefore, actual gains would have been underestimated in our evaluation. However, these underestimates did not have adverse effects on the results we obtained from our evaluation because estimated gain was only used for predicting the appropriate threshold of commonality. Because we did not apply statistical tests, threats to conclusion validity still remained.

VI. Conclusion

We proposed and evaluated a method of deriving threat pairs from several STs for security requirements elicitation. One threat in an ST in the method is related to another threat in another ST according to SFR-commonality, which is calculated on the basis of SFRs used in both threats. Several STs are thus integrated together (response to RQ1). Threat pairs contribute to expanding the candidates for security requirements when a security requirement against a threat in an ST has already been found. Because SFR does not depend on languages such as English or French, we can integrate STs written in any languages (response to RQ2). We need a threshold to determine whether two threats with a certain SFR-commonality may really be examined together as candidates for security requirements. The threshold can be decided by observing the change of estimated gain, which could be calculated without knowledge as to whether two threats really threaten the same requirement respectively (response to RQ3).

TABLE III. STS IN MULTI-FUNCTION DEVICE (MFD) DOMAIN.

ID	No.# of Threats	Summary
C291	2	Toshiba Tec, e-Studio 555/655/755/855, Jun. 2009, in Japanese
C272	2	Kyocera, Data Security Kit (E) Software Type IV ver. 1.10, Aug. 2010, in Japanese
C281	5	Sharp, MX-FR22 ver. 0.05, Jul. 2011, in Japanese

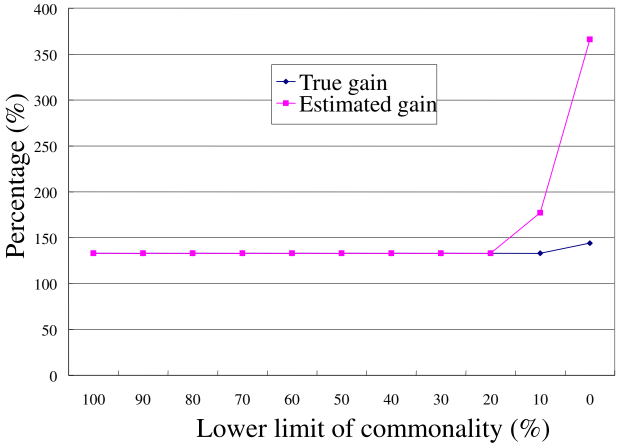


Figure 9. Changes in true and estimated gain of threat pairs in MFD domain.

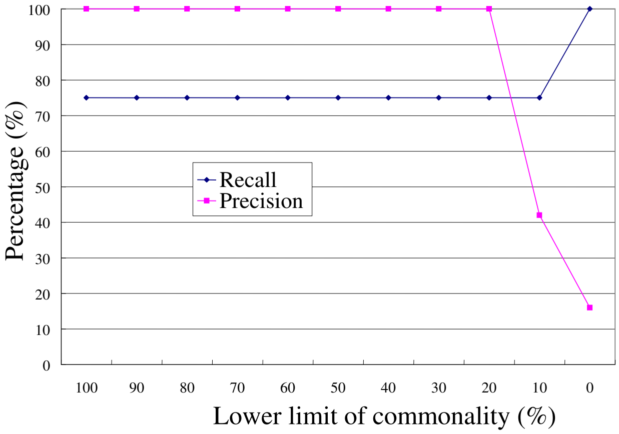


Figure 10. Changes in recall and precision of threat pairs in MFD domain.

There are a great deal of security knowledge in the field of security [18], and some of it is highly structured, is easy for computers to handle, and is usually up to date. For example, Common Attack Patterns Enumeration and Classification (CAPEC) [8] provides knowledge on attacks with several useful attributes such as attack prerequisites (preconditions of an attack) and typical ways of mitigation although this knowledge is mainly about design or implementation issues. Knowledge is also provided in XML documents. We want to extend our method so that it can be applied to various kinds of structured and machine-readable security documents.

ACKNOWLEDGMENT

This work was supported by KAKENHI 15H02686.

REFERENCES

[1] K. Breitman and J. C. S. do Prado Leite, “Ontology as a requirements engineering product,” in RE, 2003, pp. 309–319.

[2] S. W. Lee and R. A. Gandhi, “Ontology-based active requirements engineering framework,” in APSEC, 2005, pp. 481–490.

[3] H. Kaiya and M. Saeki, “Using domain ontology as domain knowledge for requirements elicitation,” in RE, 2006, pp. 186–195.

[4] D. V. Dzung and A. Ohnishi, “Improvement of quality of software requirements with requirements ontology,” in QSIC, 2009, pp. 284–289.

[5] M. Kitamura, R. Hasegawa, H. Kaiya, and M. Saeki, “An integrated tool for supporting ontology driven requirements elicitation,” in ICSOFT (SE), 2007, pp. 73–80.

[6] I. Omoronyia, G. Sindre, T. Stålhane, S. Biffl, T. Moser, and W. D. Sunindyo, “A domain ontology building process for guiding requirements elicitation,” in REFSQ, 2010, pp. 188–202.

[7] “Certified Products : New CC Portal,” http://www.commoncriteriaportal.org/products/ [accessed: 2015-08-19].

[8] “CAPEC - Common Attack Pattern Enumeration and Classification (CAPEC),” http://capec.mitre.org/ [accessed: 2015-08-19].

[9] M. Saeki, S. Hayashi, and H. Kaiya, “Enhancing Goal-Oriented Security Requirements Analysis Using Common Criteria-Based Knowledge,” International Journal of Software Engineering and Knowledge Engineering, vol. 23, no. 5, Jun. 2013, pp. 695–720.

[10] Y. Zhao, J. Dong, and T. Peng, “Ontology classification for semantic-web-based software engineering,” IEEE T. Services Computing, vol. 2, no. 4, 2009, pp. 303–317.

[11] F. Massacci, J. Mylopoulos, F. Paci, T. T. Tun, and Y. Yu, “An extended ontology for security requirements,” in Advanced Information Systems Engineering Workshops - CAiSE 2011 International Workshops, London, UK, June 20-24, 2011. Proceedings, 2011, pp. 622–636.

[12] A. Souag, C. Salinesi, R. Mazo, and I. Comyn-Wattiau, “A security ontology for security requirements elicitation,” in Engineering Secure Software and Systems - 7th International Symposium, ESSoS 2015, Milan, Italy, March 4-6, 2015. Proceedings, 2015, pp. 157–177.

[13] J. S. Dong, Y. Feng, Y. F. Li, and J. Sun, “A tools environment for developing and reasoning about ontologies,” in APSEC, 2005, pp. 465–472.

[14] A. Zouaq and R. Nkambou, “Evaluating the generation of domain ontologies in the knowledge puzzle project,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 11, 2009, pp. 1559–1572.

[15] M. Li and F. Zang, “A self-feedback methodology of domain ontology modeling,” Software Engineering, World Congress on, vol. 2, 2009, pp. 218–223.

[16] K. K. Breitman and J. C. S. do Prado Leite, “Lexicon based ontology construction,” in SELMAS, 2003, pp. 19–34.

[17] M. Karyda, T. Balopoulos, L. Gymnopoulos, S. Kokolakis, C. Lambrinoudakis, S. Gritzalis, and S. Dritsas, “An ontology for secure e-government applications,” in Proceedings of the The First International Conference on Availability, Reliability and Security, ARES 2006, The International Dependability Conference - Bridging Theory and Practice, April 20-22 2006, Vienna University of Technology, Austria, 2006, pp. 1033–1037.

[18] A. Hazeyama, “Survey on body of knowledge regarding software security,” in SNPD, 2012, pp. 536–541.

[19] M. Saeki and H. Kaiya, “Security requirements elicitation using method weaving and common criteria,” in MoDELS Workshops, 2008, pp. 185–196.

[20] K. Taguchi, N. Yoshioka, T. Tobita, and H. Kaneko, “Aligning security requirements and security assurance using the common criteria,” in

Proc. 4th International Conference on Secure Software Integration and Reliability Improvement (SSIRI'10), 2010, pp. 69–77.

[21] M. Ware, J. Bowles, and C. Eastman, "Using the common criteria to elicit security requirements with use cases," in Proc. IEEE Southeast Conference, 2005, pp. 273–278.

[22] P. Jaccard, "The distribution of the flora in the alpine zone," New Phytologist, vol. 11, 1912, pp. 37–50, doi: 10.1111/j.1469-8137.1912.tb05611.x.

[23] International Standard, " ISO/IEC 27002 Information technology - Security techniques - Code of practice for information security management," Jun. 2005.

Using ODA Method and FOIL Algorithm to Determine Organizational Agility Level

Gusts Linkevics, Uldis Sukovskis
Institute of Applied Computer Systems
Riga Technical University
Riga, Latvia
e-mail: gusts.linkevics@rtu.lv, uldis.sukovskis@rtu.lv

***Abstract* — Agile software development has become more common during the past decade. Transitioning an organization to be agile is not an undemanding task, and it requires an active involvement of all the stakeholders. The main issue is that organizations are not aware of their agility level and the necessary operations they should perform to become more agile. The purpose of this research is to use Agility Impact Index (AII) in combination with Organization Agility Model (OAM), question generation algorithm and First-Order Inductive Learner (FOIL) algorithm to calculate the agility level of an organization. The result of the research is a proof of the Organization Domain Agility (ODA) method concept. The intention of the ODA method is to determine the agility level of a Software Development Company (SDC), which is an important step to improve the agility level of an organization.**

Keywords - Agile; Software development; FOIL; AII.

I. Introduction

Agile software development is being implemented by an increasing number of Software Development Companies (SDCs). SDCs use various agile methods, for example, Extreme Programming or Scrum [3]. There are two common issues that SDCs encounter during agile method application. One of the issues is related to the lack of awareness about their efficiency in applying agile methodology. The other common issue is the incapability to identify the exact problem in the agile methodology implementation process. Such problems can be solved by hiring agile experts or by training internal employees, which may require a significant amount of time and financial resources. The other approach is to use a method which helps to solve this problem by determining the problematic areas. Organization Domain Agility (ODA) [1] is one of the methods used for a periodical evaluation of the organisation and the team to determine the problematic areas.

The main focus of the paper is to provide a compendious introduction of the ODA method and to present detailed findings about its main components which enable to determine the agility level of an organization:

- Agility Impact Index (AII);
- Question generation algorithm;
- Domain, Sub-domain and attribute Value Tree (DSA Value Tree);
- First level rule generation using FOIL [6] algorithm.

This paper consists of 6 sections. In Section 1 the problem of the organization agility and the goal of the paper is described. In Section 2 the ODA method and its process is introduced. The ODA method is developed to evaluate the agility level of SDCs. In Section 3 the Question generation process is explained. Generated questions are used to gather the data about SDCs. In Section 4, the DSA Value Tree is depicted, and it consists of AII values. AII value is defined for each element of the DSA tree. AII values are determined by the group of experts. In Section 5, we focus on FOIL algorithm usage for the evaluation of the SDC agility level. Section 6 concludes the paper and provides an outline of the future work.

II. ODA Method

ODA method uses Organization Agility Model (OAM) [1] to describe an organization and its team in a structured way. The structural approach provides an opportunity to evaluate various parts of the organization. OAM is organized in a tree structure where the organization is described by the DSA Value Tree. The DSA Value Tree groups similar items of the organization into domains and sub-domains.

The initial model consisted of five domains [2]: Organization, Productivity, Process, Quality and Value. During the more detailed research it was noticed that an additional top level domain is needed to describe the project component of the organization (Figure 1).

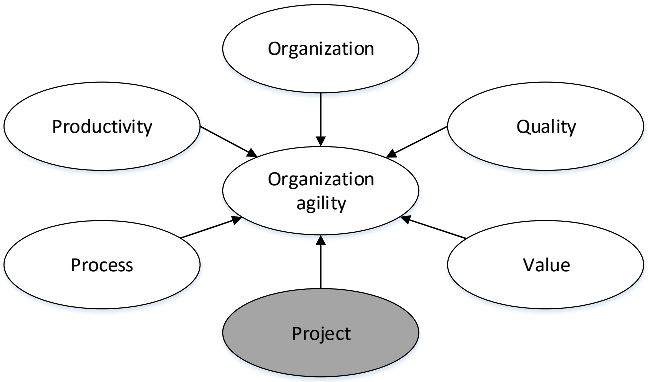


Figure 1. Extended OAM .

The purpose of the Project domain is to describe attributes of the particular project. Different projects in the SDC can be at different agility levels and can influence the Organizational agility differently.

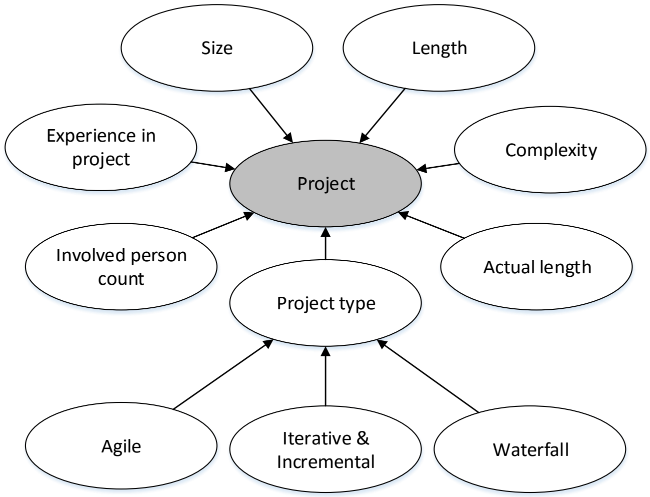


Figure 2. Project Domain Components.

Figure 2 shows the structure of the Project domain, and it consists of seven sub-domains:

- Project type – it is possible to identify 3 project subtypes:
  - Waterfall – this is a classical approach to software development.
  - Iterative and Incremental – this approach is a custom type of Agile methodology.
  - Agile – a project is based on Scrum or a similar method.
- Number of people involved – agile methods work better with small number of people within a team. In case of large projects, there is some overhead in managing the large teams, and it can decrease the overall agility level of the project.
- Experience in a project – experience of the team with the particular project also influences the agility level of the team and organization. In case the project has been running for several years and it is necessary to switch it to the agile development approach, some experienced employees may resist using the agile method in the project.
- Size – smaller projects are easier to shift to the agile approach than larger projects.
- Length – similarly to the size of the project the length of the project also will influence the agility level of the organization.
- Complexity – sometimes it is related to the project size. Complex projects fail more frequently than less complex projects as the complex projects require more formal approach. The approach could still be agile, but the more complex projects require more detailed documentation.
- Actual length (Age) – the time the project has been already running. As mentioned before, projects running for a long period of time have a potentially higher risk when switching to the agile approach than starting a new project completely with agile from the beginning.

There are several attributes that describe each sub-domain. The attribute values for the Complexity sub-domain are shown in TABLE I, and the Size (Amount of the investment) sub-domain attributes are listed in TABLE II. The list of all attributes is not included in the paper due to the limited space.

TABLE I. COMPLEXITY SUB-DOMAIN ATTRIBUTES

Attribute	Description
Low	Project is simple and does not have complex integrations and components.
Medium	Project has some integrations, but they are not complex. Project has several components which need to be integrated.
High	Complex algorithms, integrations and components are used.

TABLE II. SIZE SUB-DOMAIN ATTRIBUTES

Attribute	Investment amount
Enhancement	x < $250,000
Small	$250,000 < x < $ 1M
Medium	$ 1M < x < $ 3M
Large	$ 3M < x < $ 10M
Very large	x > $ 10M

AII values range from 1 to 10, where 1 means that the DSA item does not influence organization agility, and 10 means that the item significantly impacts the agility, for example, the Productivity domain does not influence the agility level the same way as Organization domain. AII is determined by the expert evaluation method DELPHI [5] which uses an external agile expert network to evaluate the common DSA. Agile experts do not need any information about the particular organization, and they are not directly related to it. The evaluation they provide is bound to the common DSA, which is then used together with the information acquired from the particular organization. In general, there is a basic agile knowledge which can be applied to any organization looking towards agile software development.

Agile experts evaluate the DSA at least once and then repeat it if the structure of the DSA is changed. It is possible to change the DSA structure for the ODA method (Figure 3.) in case the organization uses any other agile method than Scrum.

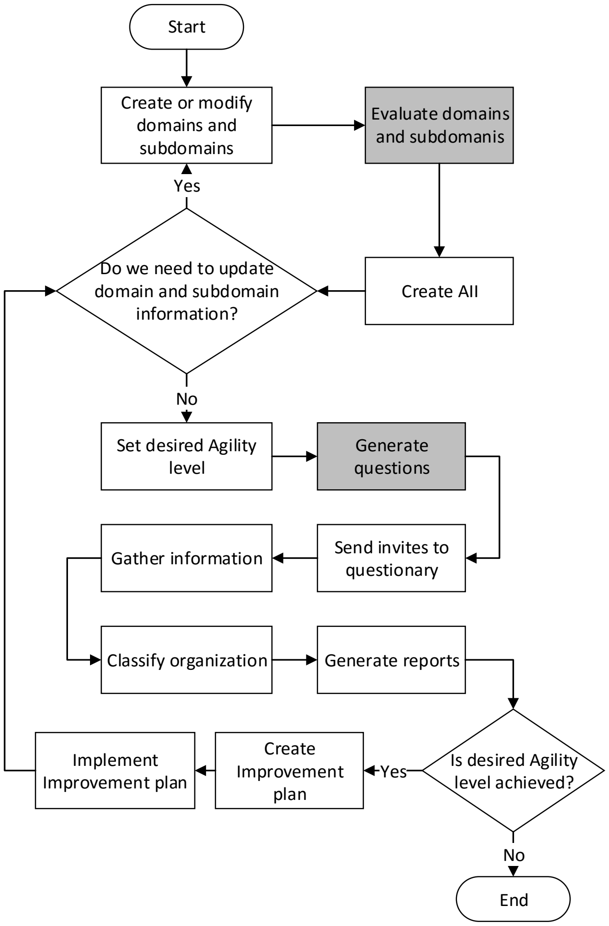

Start
Create or modify domains and subdomains
Evaluate domains and subdomanis
Yes
Do we need to update domain and subdomain information?
Create AII
No
Set desired Agility level
Generate questions
Gather information
Send invites to questionary
Classify organization
Generate reports
Implement Improvement plan
Create Improvement plan
Yes
Is desired Agility level achieved?
No
End

Figure 3. Process of a Method ODA.

After all of the AII values for the DSA are determined, the Question generation algorithm is used to generate question sets for the employees. In the next section, the Question generation algorithm is described in more detail. The generated questions are distributed among employees in different departments.

III. QUESTION GENERATION

Question generation is an important part of the ODA method, and it is required to generate only a small set of questions for each employee at each evaluation period. There are approximately 300 questions to ask, and it is impractical to ask each employee all of them. Based on the question ranking by AII value, the most influential questions are asked first.

As it is shown in Figure 4 the Question generator generates subsets of questions from the set of all questions (1).

$$Q = \{q_1, q_2, q_3 \dots q_m\} \qquad (1)$$

Where:

- Q – Set of all questions.
- $q_{1\dots m}$ – Questions, where m is the total number of questions.

An employee based question set can be defined as a subset of all questions (2).

$$A_{1\dots n} \in Q \qquad (2)$$

Where:

- Q – Set of all questions.
- $A_{1\dots n}$ – Subsets of questions for an employee, where n is the amount of employees participating in survey.

The organization sets the number of questions for each employee. It is not recommended to create large sets of questions as it may lead to low quality of answers. It is recommended to include up to 10 questions in each evaluation [8], and the evaluation process should take from 5 to 7 minutes.

There are three types of questions in the employee question set (3):

- Priority questions – the initiator of question generation marks a number of questions to be included in all of the generated question sets. The priority questions make 20 per cent of all the questions in the set.
- Unanswered questions ordered by AII – a list of all the unanswered questions ordered by AII value. After adding High priority questions to the set, the unanswered questions are added to the set. This is required to cover the maximum amount of information about the DSA. At the beginning the most influential questions are added. Those questions make 60 per cent of all the questions in the set.
- Previously answered questions ordered by AII – this type of questions helps to keep the “pulse” on the most influential DSA elements, and those questions form 20 per cent of all the questions.

$$A_{1\dots n} = \{P_q, N_{1\dots n}, O_{1\dots n}\}. \qquad (3)$$

Where:

- $A_{1\dots n}$ – Question set for particular employee.
- P_q – Set of priority questions.
- $N_{1\dots n}$ – Unanswered questions ordered by AII.
- $O_{1\dots n}$ – Previously answered questions ordered by AII.

The question generation process is shown in Figure 4. For example, if there are 17 questions in the question set Q, as in

$$Q = \{q_1=9,\ q_2=8,\ q_3=7,\ q_4=6,\ q_5=9,\ q_6=8,\ q_7=7,\ q_8=7,\ q_9=8,\\ q_{10}=9,\ q_{11}=9,\ q_{12}=8,\ q_{13}=7,\ q_{14}=7,\ q_{15}=6,\ q_{16}=7,\ q_{17}=8\}. \qquad (4)$$

And 4 questions in priority question set P_q (Priority questions have been selected by generation initiator) as in

$$P_q = \{q_{10}=9,\ q_{17}=8,\ q_3=7,\ q_{15}=6\}. \qquad (5)$$

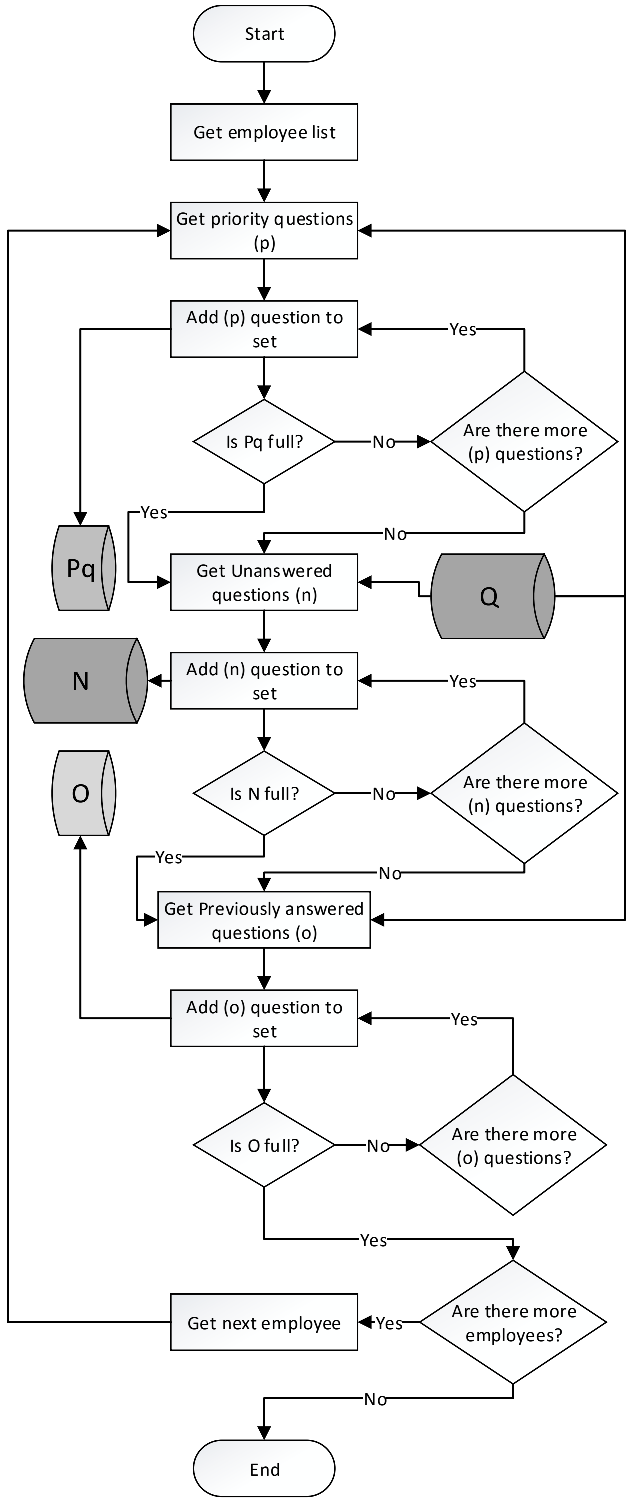

Start
Get employee list
Get priority questions (p)
Add (p) question to set
Yes
Is Pq full?
No
Are there more (p) questions?
Yes
No
Pq
Get Unanswered questions (n)
Q
N
Add (n) question to set
Yes
Is N full?
No
Are there more (n) questions?
O
Yes
No
Get Previously answered questions (o)
Add (o) question to set
Yes
Is O full?
No
Are there more (o) questions?
Yes
Get next employee
Yes
Are there more employees?
No
End

Figure 4. Question Generation Process.

And 4 questions in the previously answered question set O, as in

$$O = \{q_{12}=8, q_7=7, q16=7, q_4=6\}. \quad (6)$$

And there is one employee involved in the evaluation of 10 questions, then the resulting question set would contain questions

$$A_1 = \{P\{q10,q17\},N\{q1,q5,q11,q2,q6,q9\},O\{q12,q7\}\}. \quad (7)$$

Depending on the number of employees correct timing for question generation should be selected. As question generation depends heavily on question sets N and O, then it is reasonable to assume that the questions are generated during the night. In this way, also the system is not congested during working hours.

After question generation, question sets are sent to each employee. Time for question sending should be selected properly [9], in this case, after the Review meeting and before the Retrospective meeting. The gathered information is used to build the DSA Value Tree.

IV. DSA VALUE TREE

DSA value tree is a way to represent the gathered data. Tree view is a convenient way to identify the problematic areas and compare the gathered data with the AII values identified by the expert.

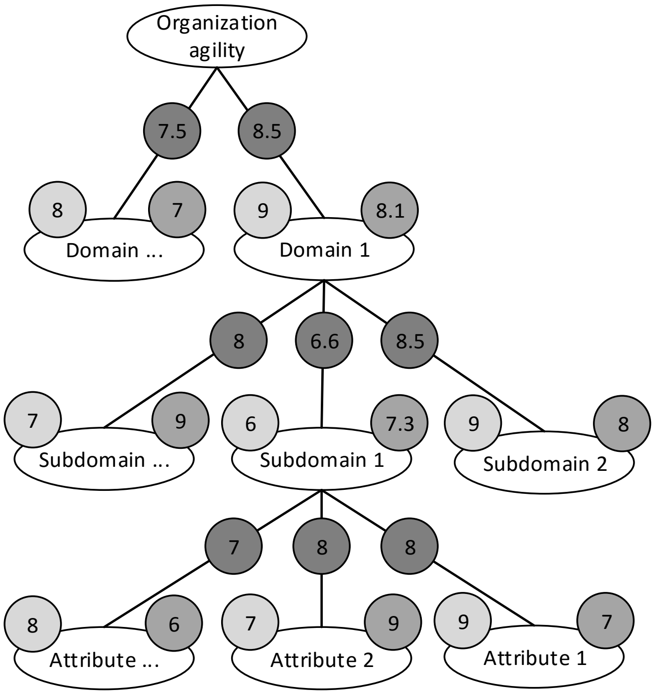


Figure 5. DSA value tree example.

Figure 5 shows a sample of a DSA Value Tree where the values on the right are the values gathered from the employee surveys (ESV), and the values on the left are the AII values defined by the experts. ESV values are calculated using average weighted values (5).

$$\bar{y} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad (8)$$

Where:

- $\bar{y}$ – Set of all questions.
- w_i – weight value, in this case AII.
- x_i – ESV value, determined from surveys.
- n – Number of respondents who have answered particular question.

The agility level can be determined on an organization level, on a project level or on a team level. The following breakdown is required to improve the agility level at a particular team or a project. As mentioned before, the agility level can differ on a team or a project level. In case of a team or a project agility level determination, filtering of set $\bar{y}$ is used to include only ESV values for a particular team or a project (Figure. 6).

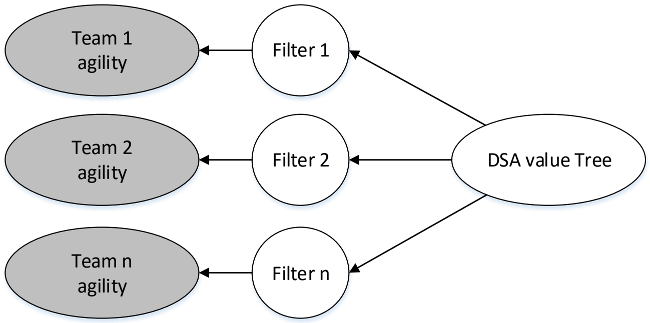


Figure 6. DSA Value Tree filtered for the team.

One of the ways to represent the gathered data is grouping the data by project and team after the DSA value tree is created (Figure. 7).

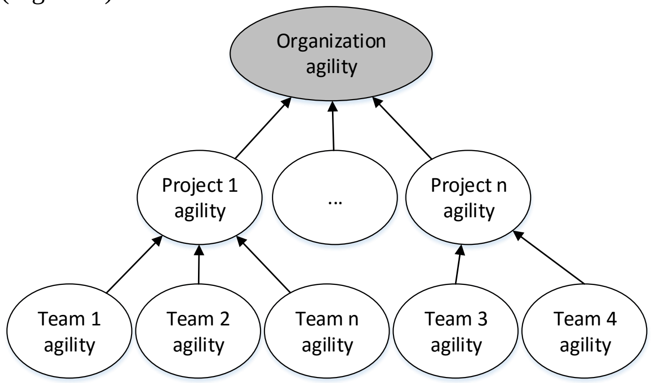


Figure 7. Grouping of the DSA value tree.

The grouping approach helps to identify the problematic projects and teams where additional effort for the agility level increment is needed.

After all $\bar{y}$ values are calculated, the rules based on these values can be created. The ODA method uses FOIL algorithm to generate the rules for the final evaluation of the SDC agility level.

V. Using FOIL Algorithm For The Rule Generation

FOIL is a system for finding function-free Horn clauses [6]. FOIL searches for the first-order rules using a learning set. The search results in finding a set of logical rules describing the system under consideration.

The first-order rule is a logical proposition of the form:

$$R(V1,V2,\ldots,Vk) \leftarrow L_1, L_2, \ldots, L_m, \qquad (9)$$

Where:

- R is a target relation between variables V_i.
- L_i are literals composing a condition which verity enables one to state that the head of the rule is true.

Within this research FOIL is used to determine the agility level of the SDC where the set of the first-order rules determines the agility class of the SDC. Before generating the first-order rules for determining the agility level of an Organization, a Project or a Team, it is necessary to define the agility classes. The Agility class determines the present agility level of an organization, a project or a team. Knowledge helps to create a specific improvement plan and to implement it later. It is reasonable to use 5 or 10 agility classes of evaluation, as it is with grades at school. Some school systems use 5 point grading system, whereas some schools use 10 point grading system. It depends on how accurately we want to evaluate. In this case, it is decided to use 5 agility classes (K1, K2, K3, K4 and K5) which will map to the average values of the DSA Value Tree. Each class corresponds to 2 values (Table III).

TABLE III. Agility Class Mapping to the DSA Value Tree

DSA Value Tree values	Agility Class	Description
1, 2	K1	Not agile and no evidence of agility
3, 4	K2	Not agile, but some evidence of agility exists
5, 6	K3	Some evidence of agility, but major improvements should be introduced
7, 8	K4	Agile, but some problems exist and requires some improvements
9, 10	K5	Agile and no important improvements are needed.

The first-order rules are important for the agility class determination because it is not possible to determine the exact level of agility there is only the information about the DSA Value Tree. For example, if the top-level domain average values are 1, 4, 5, 6, 7 and 8 (each value represents average value of the Organization, Productivity, Quality, Project, Value and Process domain), one should analyse in more detail if it means the organization is agile. The same question may be discussed if the DSA Value Tree values are 5, 2, 7, 9, 3 and 8. The FOIL algorithm can be used to resolve such problems.

FOIL algorithm uses learning data set which contains information about the known organisations valued by experts. The learning data set has information about the specific organisation values and its class. There is a small learning data sample presented in Table IV. The data listed in the Table IV has been simplified to shorten the solution (Value ranges are shortened and not all the domains are included). Columns D1, D2, D3 and D4 represent four top level domains Organization, Process, Productivity and Quality. The sample learning data set contains learning data only for three agility classes N1 (Not agile and no signs of agility), N2 (Some signs of agility, but major improvements should be introduced) and N3 (Agile and no important improvements are needed) and is also simplified.

TABLE IV. SAMPLE FOIL LEARNING DATA

D1	D2	D3	D4	Class
1	1	1	1	N1
1	1	1	2	N2
1	1	2	1	N1
1	1	2	2	N3
1	2	1	1	N1
1	2	1	2	N2
1	2	2	1	N1
1	2	2	2	N3
3	1	1	1	N1
3	1	1	2	N2
3	1	2	1	N1
3	1	2	2	N3
3	2	1	1	N1
3	2	1	2	N2
3	2	2	1	N1
3	2	2	2	N1
2	1	1	1	N1
2	1	1	2	N1
2	1	2	1	N1
2	1	2	2	N3
2	2	1	1	N1
2	2	1	2	N2
2	2	2	1	N1
2	2	2	2	N1

First-order rules can be described in a form of IF … THEN … or in a form of Horn clauses, Head ← Body [6]. In this case Agility Class ← Condition. In case of simplified learning data there are three agility classes N1, N2 and N3. Each class in the learning data set has a specific number of records N1 = 15, N2 = 5 and N3 = 4. The learning data set is used to teach the algorithm how to identify particular agility class.

FOIL algorithm uses FOIL_GAIN function to evaluate each next literal to be added to the class identification rule (10) [6].

$$Foil\_Gain(L, R) \equiv t \left(\log_2 \frac{p_1}{p_1 + n_1} - \log_2 \frac{p_0}{p_0 + n_0} \right) \quad (10)$$

Where:

- L – New condition to be added to the rule.
- R – Rule body, to which we want to add the condition.
- p_0 – Number of positive items in rule R.
- n_0 – Number of negative items in rule R.
- p_1 – Number of positive items in rule R_1.
- n_1 – Number of negative items in rule R_1.
- T – Number of positive items in rule R after adding new condition L to the rule R.

To create rules for the class N1 we need to identify all positive and negative examples, as seen in Table V and Table VI.

TABLE V. POSITIVE SAMPLES FOR CLASS N1

D1	D2	D3	D4	Class
1	1	1	1	N1
1	1	2	1	N1
1	2	1	1	N1
1	2	2	1	N1
3	1	1	1	N1
3	1	2	1	N1
3	2	1	1	N1
3	2	2	1	N1
3	2	2	2	N1
2	1	1	1	N1
2	1	1	2	N1
2	1	2	1	N1
2	2	1	1	N1
2	2	2	1	N1
2	2	2	2	N1

TABLE VI. NEGATIVE SAMPLES FOR CLASS N1

D1	D2	D3	D4	Class
1	1	1	2	N2
1	1	2	2	N3
1	2	1	2	N2
1	2	2	2	N3
3	1	1	2	N2
3	1	2	2	N3
3	2	1	2	N2
2	1	2	2	N3
2	2	1	2	N2

During the next step a new literal should be added (11).

$$I(T_1) = -\log_2(15/(15+9)) = 0{,}678 \quad (11)$$

After checking the results (Table VII) a literal D4(X, 1) can be added to the rule N1 ← D4(X, 1).

TABLE VII. CALCULATION OF FOIL_GAIN FOR CLASS N1

Literal	Calculation	Foil_Gain
D4(X, 1)	I(T_2) = -log$_2$(12/(12+0)) = -log$_2$(1)=0	12*(0,678-0) = 8,136
D4(X, 2)	I(T_2) = -log$_2$(3/(3+9)) = -log$_2$(0,25)=2	3*(0,678-2) = -3,966
D3(X, 2)	I(T_2) = -log$_2$(8/(8+4)) = -log$_2$(0,666)= 0,586	8*(0,678-0,586) = 0,736
D3(X, 1)	I(T_2) = -log$_2$(7/(7+5)) = -log$_2$(0,583)= 0,788	7*(0,678-0,788)=- 0,77
D2(X, 1)	I(T_2) = -log$_2$(7/(7+5)) = -log$_2$(0,583)= 0,788	7*(0,678-0,788)=- 0,77
D2(X, 2)	I(T_2) = -log$_2$(8/(8+4)) = -log$_2$(0,666)= 0,586	8*(0,678-0,586) = 0,736
D1(X,1)	I(T_2) = -log$_2$(4/(4+4)) = -log$_2$(0,5)= 1	4*(0,678-1) = -1,288
D1(X,3)	I(T_2) = -log$_2$(5/(5+3)) =	5*(0,678-0,678) = 0

Literal	Calculation	Foil_Gain
	-log$_2$(0,625)= 0,678	
D1(X,1)	I(T_2) = -log$_2$(6/(6+2)) = -log$_2$(0,75)= 0,415	2*(0,678-0,415) = 0,263

As this condition does not return any negative samples, it is possible to stop the further processing of the rule. However, as this rule does not return all the positive samples, it is necessary to add an additional condition to the rule set for class N1. After removing all the positive samples covered by the first condition, there are only 3 positive samples left (Table VIII).

TABLE VIII. POSITIVE SAMPLES FOR CLASS N1, AFTER REMOVING POSITIVE SAMPLES AFFECTED BY FIRST CONDITION

D1	D2	D3	D4	Class
3	2	2	2	N1
2	1	1	2	N1
2	2	2	2	N1

During the following step the next literal should be added (12).

$$I(T_1) = -\log_2(3/(3+9)) = 2 \qquad (12)$$

TABLE IX. CALCULATION OF FOIL_GAIN FOR CLASS N1

Literal	Calculation	Foil_Gain
D4(X, 2)	I(T_2) = -log$_2$(3/(3+9)) = -log$_2$(0,25)=2	3*(2-2) = 0
D3(X, 2)	I(T_2) = -log$_2$(2/(2+4)) = -log$_2$(0,333)= 1,586	2*(2-1,586) = 0,414
D3(X, 1)	I(T_2) = -log$_2$(1/(1+5)) = -log$_2$(0,166)= 2,59	1*(2-2,59) = -0,59
D2(X, 1)	I(T_2) = -log$_2$(1/(1+5)) = -log$_2$(0,166)= 2,59	1*(2-2,59) = -0,59
D2(X, 2)	I(T_2) = -log$_2$(2/(2+4)) = -log$_2$(0,333)= 1,586	2*(2-1,586) = 0,414
D1(X,1)	I(T_2) = -log$_2$(0/(0+4)) = -log$_2$(0)= ∞	-
D1(X,3)	I(T_2) = -log$_2$(1/(1+3)) = -log$_2$(0,25)= 2	1*(2-2) = 0
D1(X,2)	I(T_2) = -log$_2$(2/(2+2)) = -log$_2$(0,5)= 1	2*(2-1) = 2

As shown in Table IX, the most notable gain is from literal D1(X, 2), which can be added to the rule, but, as it also selects negative samples, more literals should be added.

TABLE X. CALCULATION OF FOIL_GAIN FOR CLASS N1

Literal	Calculation	Foil_Gain
D4(X, 2)	I(T_2) = -log$_2$(2/(2+2)) = -log$_2$(0,5)=1	2*(2-1) = 2
D3(X, 2)	I(T_2) = -log$_2$(1/(1+1)) = -log$_2$(0,5)=1	1*(2-1) = 1
D3(X, 1)	I(T_2) = -log$_2$(1/(1+1)) = -log$_2$(0,5)=1	1*(2-1) = 1
D2(X, 1)	I(T_2) = -log$_2$(1/(1+1)) = -log$_2$(0,5)=1	1*(2-1) = 1
D2(X, 2)	I(T_2) = -log$_2$(1/(1+1)) = -log$_2$(0,5)=1	1*(2-1) = 1

To this point the second rule is N1 ← D1(X, 2) ∧ D4(X, 2) and it has to be checked if it returns any negative samples. Considering the fact that it returns negative samples, the additional literal should be added.

TABLE XI. CALCULATION OF FOIL_GAIN FOR CLASS N1

Literal	Calculation	Foil_Gain
D3(X, 2)	I(T_2) = -log$_2$(1/(1+1)) = -log$_2$(0,5)=1	1*(2-1) = 1
D3(X, 1)	I(T_2) = -log$_2$(1/(1+1)) = -log$_2$(0,5)=1	1*(2-1) = 1
D2(X, 1)	I(T_2) = -log$_2$(1/(1+1)) = -log$_2$(0,5)=1	1*(2-1) = 1
D2(X, 2)	I(T_2) = -log$_2$(1/(1+1)) = -log$_2$(0,5)=1	1*(2-1) = 1
D3(X, 2)	I(T_2) = -log$_2$(1/(1+1)) = -log$_2$(0,5)=1	1*(2-1) = 1

As shown in Table XI, in this case all the literals are equally bad as all return negative samples. As a consequence, one more literal should be added to the rule. To this point the rule is N1 ← D1(X, 2) ∧ D4(X, 2) ∧ D3(X, 2).

TABLE XII. CALCULATION OF FOIL_GAIN FOR CLASS N1

Literal	Calculation	Foil_Gain
D2(X, 1)	I(T_2) = -log$_2$(0/(0+1)) = -log$_2$(0)=∞	-
D2(X, 2)	I(T_2) = -log$_2$(1/(1+0)) = -log$_2$(1)=1	1*(2-0) = 2

At this point literal D2(X, 2) can be added to the rule set N1 ← D1(X, 2) ∧ D4(X, 2) ∧ D3(X, 2) ∧ D2(X, 2).

Acknowledging the fact that all the positive examples are not yet covered, an additional rule should be added to the set. All the positive samples covered by the new rule should be removed from the learning set, and a search of a new rule should be continued. The process is repeated as many times as necessary until all the positive samples are covered. At the end, the rule set for class N1 looks like:

- N1 ← D4(X, 1)
- N1 ← D1(X, 2) ∧ D4(X, 2) ∧ D3(X, 2) ∧ D2(X, 2)
- N1 ← D1(X, 2) ∧ D4(X, 2) ∧ D3(X, 1) ∧ D2(X, 1)
- N1 ← D1(X, 3) ∧ D3(X, 2) ∧ D2(X, 2)

This process is repeated for class N2 and N3. A complete rule set for all tree classes is shown in Table XIII.

TABLE XIII. FINAL RULE SET FOR CLASSES N1, N2 AND N3

Class	Rule set

N1	N1 ← D4(X, 1)
	N1 ← D1(X, 2) ∧ D4(X, 2) ∧ D3(X, 2) ∧ D2(X, 2)
	N1 ← D1(X, 2) ∧ D4(X, 2) ∧ D3(X, 1) ∧ D2(X, 1)
	N1 ← D1(X, 3) ∧ D3(X, 2) ∧ D2(X, 2)

N2	N2 ← D4 (X, 2) ∧ D3 (X, 1) ∧ D2 (X, 2)
	N_2 ← D4 (X, 2) ∧ D3 (X,1) ∧ D2 (X, 1) ∧ D1 (X, 1)
	N2 ← D3 (X, 1) ∧ D4 (X, 2) ∧ D1 (X, 3)

N3	N_3 ← D4 (X, 2) ∧ D3 (X, 2) ∧ D1 (X, 1) N_3 ← D4 (X, 2) ∧ D3 (X, 2) ∧ D2 (X, 1)

When the learning data set is processed, the rule sets can be tested against the learning data set (Table XIV). In this case, rules generated by FOIL can be used to identify tree classes of agility N1, N2 and N3. These rules are simplified, but the approach can be used to generate more complex rules to satisfy the needs of the ODA method.

TABLE XIV. RULE SET TESTING RESULTS

No.	D1	D2	D3	D4	Class	Result	Rule
1	1	1	1	1	N1	N1	1
2	1	1	1	2	N2	N2	6
3	1	1	2	1	N1	N1	1
4	1	1	2	2	N3	N3	8,9
5	1	2	1	1	N1	N1	1
6	1	2	1	2	N2	N2	5
7	1	2	2	1	N1	N1	1
8	1	2	2	2	N3	N3	8
9	3	1	1	1	N1	N1	1
10	3	1	1	2	N2	N2	7
11	3	1	2	1	N1	N1	1
12	3	1	2	2	N3	N3	9
13	3	2	1	1	N1	N1	1
14	3	2	1	2	N2	N2	5,7
15	3	2	2	1	N1	N1	1,4
16	3	2	2	2	N1	N1	4
17	2	1	1	1	N1	N1	1
18	2	1	1	2	N1	N1	3
19	2	1	2	1	N1	N1	1
20	2	1	2	2	N3	N3	9
21	2	2	1	1	N1	N1	1
22	2	2	1	2	N2	N2	5
23	2	2	2	1	N1	N1	1
24	2	2	2	2	N1	N1	2

One of the problems of this approach is that there is a need for a set of quality training data for the algorithm, which is not so easy to gather.

VI. CONCLUSION

Shifting a project to agile software development is not an effortless procedure, and there are different ways to accomplish it. Some organizations hire expensive agile experts, while others try to execute the transition process themselves. The ODA method can support the transition process. It has several steps, and it starts with the creation of OAM. The next step is the DSA evaluation carried by agile experts using the DELPHI method. The evaluated AII values are later used to generate the employee-based question sets. The data gathered from the questionnaires is used to create the DSA value tree. When the DSA value tree is created, it is used by the FOIL method to generate rules for determining the agility level.

The process of assessing the actual organization agility is long, but in most cases it can be completely automated. For example, the AII values are already defined for the DSA, and the organization does not need to hire any agile experts. It could be required only in cases when the existing DSA does not match the organization, especially in the process domain part. The initial process domain of the DSA is built based on Scrum, which resulted to be the most common agile method during the last few years.

As there are approximately 300 questions in the question set Q, there is a risk to fail in collecting the necessary data from all the employees. To mitigate this risk, the Question generation algorithm is used to generate smaller subsets of questions each time. The Question set size depends on the SDC. Some organizations could generate sets of 10 questions whereas other organizations could generate 15 questions per set. The question amount per set is configurable in the supporting tool of the ODA method. The Question sets are generated periodically, and they include three types of questions. There are High priority questions which are included in the question set if it is necessary to gather the feedback within a short period of time. There are Unanswered questions ordered by AII and Answered questions with high AII values which need to be answered more frequently.

There is additional risk related to AII values. In order to make ODA method work correctly, the agile expert network should be of a high quality and expertise. High quality expert network creation is not an easy task, but it is achievable, and mostly the SDC who use the ODA method will not need to create the network themselves.

Rule generation using FOIL is automated, and the algorithm is suitable for grouping tasks. It is assumed that it is possible to use similar algorithms as well. The biggest challenge at this step is to have a good quality learning data for the algorithm as the quality of the generated rules depends on the quality of the learning data.

During the further research it is planned to test the method and the used algorithms on several organizations, as the concept of this approach proves to be beneficial.

REFERENCES

[1] G. Linkevics, “Evaluation of Agility in Software Development Company” Joint International Conference on Engineering Education & International Conference on Information Technology (ICEE/ICIT 2014), 2014, pp. 102.

[2] G. Linkevics, “Adopting to Agile Software Development”, Applied Computer Systems, 2014, pp. 64 -70.

[3] K. S. Rubin, “Essential Scrum: A Practical Guide to Most Popular Agile Process”, Addision-Wesley Professional, 2012.

[4] M. Poppendieck and T. Poppendieck, “Implementing Lean Software Development: From Concept to Cash”, Addision-Wesley Professional, 2006.

[5] G. J. Skulmoski, F. T. Hartman and J. Krahn, “The Delphi Method for Graduate Research”, Journal of Information Technology Education, vol. 6, 2007

[6] J. R. Quinlan. “Learning logical definitions from relations”, Boston: Kluwer Academic Publishers, 1963, pp. 239–266.

[7] Manifesto for Agile Software Development, http://agilemanifesto.org/, [retrieved: 12, 2014].

[8] SurveyMonkey, How Much Time are Respondents Willing to Spend on Your Survey? https://www.surveymonkey.com/blog/2011/02/14/survey_completion_times/, [retrieved: 8, 2015].

[9] Fluid Surveys University, It’s All About Timing –When to Send your Survey Email Invites? http://fluidsurveys.com/university/its-all-about-timing-when-to-send-your-survey-email-invites/, [retrieved: 8, 2015].

Optimising development process and software maturity through eScience partnerships

Dieter Kranzlmüller
Leibniz Supercomputing Centre (LRZ)
Garching n. Munich, Germany
Email: kranzlmueller@lrz.de

Matti Heikkurinen
MNM-Team
Ludwig-Maximilians-Universität München (LMU)
Munich, Germany
Email: heikku@nm.ifi.lmu.de

***Abstract* — Computational modelling is a crucial tool in numerous basic and applied research domains. Running the simulations at unprecedented scales on supercomputer systems represents an important catalyst for improving the performance and maturity of the software. In this paper, we present a conceptual model justifying the use of software scalability as a viable proxy indicator for the maturity of the software and its development process. We also present two approaches – workshop and partnership – that allow supercomputing centres to play a more active role as partners instead of providers during the development and improvement of modelling tools. This is confirmed by the scalability results achieved.**

Keywords: software development; computational modelling; supercomputing; research; software engineering; IT service management

I. INTRODUCTION

The process that takes algorithmic results from various basic research activities and gradually turns them into software-based, infrastructure-like services running on high-end possibly interconnected, computational systems is an important part of computational science. The requirements on the working practices in the opposite ends of this process are quite orthogonal, which may lead to poor alignment of goals and incentives between researchers and computational service providers. In the (basic) research domain, repeating the previous results achieved before in the exact same form (similar data, same tools, same environment and so on) is rarely of interest. In contrast, with *"infrastructure-like"* services uniformity, repeatability and high volume of use are indicators of value. Paradoxically, when scaling these models to high-end supercomputing environments, the work addresses both of these areas: the accuracy of the models can be verified much more reliably, while at the same time the scaling challenges expose weaknesses in the implementation, fixing of which make the software package more suitable for routine production use.

The basic research activities need to minimise the "friction" between the advances made in the theoretical explanations of the phenomena and the computational models that are used to verify theories. As a result, the first versions of the modelling software tend to be implemented by the theorists themselves, using either *ad hoc* development processes or perhaps a variation of personal agile development geared toward rapid prototyping. This imposes clear limitations on the degree of formalism that can be applied to the software development and documentation. Even if the developer of the computational models has a strong software engineering background, many of the formal methods used to improve the maturity of the software development process would have a fairly low (possibly negative) return on investment. The challenges encountered – and met – during the scalability challenges may represent one of the clearest indications that motivate developers to adapt more mature software engineering practices. For this reason, we assume that the improvements in the scalability of the software suite represent a usable proxy indicator for improvements in the development process. While obviously not 100% accurate representation of the maturity of the software development process, the advantage of this indicator is that it is possible to measure in a uniform way, independent of the specific development approach used in an individual software package.

II. CHARACTERISTICS OF THE ECOSYSTEM

A researcher working full time implementing software corresponding to the latest theoretical model might not see any short-term quality or productivity improvements through adoption of formalized, mature software development methods. It is likely that during active development phases the key parts of the software are so closely entwined with the theories being studied that (in terms of the developer him/herself) the software is in practice self-documenting. Furthermore, if the software is most likely to be discarded (due to fundamental changes in the theory being actively tested, for example) before there is a need to involve more than a handful of developers, extensive formal documentation can seem like a waste of time.

However, some of the results of such projects are retained for longer periods of time. They may also end up being relevant in more settings than initially assumed, and some of them will thus (eventually) get reused by other researchers. Often this reuse starts as a fairly informal sharing (including copy-pasting) of code between collaborating researchers, but demand for packaging the software into formal distributions is likely to emerge gradually. This will typically lead to self-organisation of the user/developer communities that formalise the

ISBN: 978-1-61208-432-9

documentation, development and distribution practices, using tools and services such as GitHub [1] or Elsevier's SoftwareX [2].

Sustaining this community-driven maturing process is often not straightforward: transitioning from the "pure" research funding to what is essentially – at least partly – product development requires adopting new business models. The ways funding of these development efforts is secured is very different compared to basic research. Instead of measuring the progress through journal impact factors and number of peer-reviewed papers, the metrics need to be linked to number of users, quality of support (with e.g. speed of closing support tickets as a proxy indicator) and – usually anecdotal – evidence of the scientific contributions enabled by the software product or service. It is exceedingly rare that the funding agency that funded the initial work and has the best understanding of the expertise of the original innovator would be able to fund both types of activities, hence a structure or organisation that acts as a mediator or fulcrum in this interplay between research and development is needed.

III. Supercomputing Centres as Fulcrums

As discussed above, the research and e-Infrastructure (or Cyberinfrastructure, as it is known in the US) funding philosophies tend to approach the "theory to computational service"-pipeline from the opposite ends. In considering the priorities of operational e-Infrastructure, the key is identifying the most often used components and measuring success – for the most part – based on the volume of use. Historically the interest (and de facto sustainable funding) has extended gradually from supporting the basic, Internet-like connectivity to more and more complex computing and data services. As an example, this development is evident when reviewing the table of contents of the White Papers of the e-Infrastructure Reflection Group [3] that have focused on areas where developments in technologies and their use cases require policy-level action. The development starting form 2003 illustrate the history of the current "layered" reference model used in the European e-Infrastructure policy work [4].

The challenge with this development path is that the maturing process starts from the opposite ends of the e-Infrastructure service and technology stack. The maturing applications should eventually "meet" the new e-Infrastructure services, but this is challenging as the developments on these two areas often proceed independently (outside specific "spearhead" applications that form the basis of the use cases supported by the new e-Infrastructure services). Without in-depth knowledge of the planned improvements of the e-Infrastructure, applications may end up with inherent limitations that prevent them from utilising all the possible benefits of the new top-level services. Similarly, the development efforts behind the standardised interfaces to provide access to supercomputing and other advanced services may not be sufficiently informed about the requirements of the existing application software solutions and their user communities. The decoupling of the typical funding sources further complicates this, as the communities lack organic meeting points (conferences, cross-project events organised by funding agencies etc.).

However, there is one area where basic research and advanced service provision meet every day: general purpose supercomputing centres. Typically such a centre supports tens or hundreds of applications and use cases, and successful day to day operations need to be based on a broad consensus on what interfaces and solutions should be supported. As an example, Leibniz Supercomputing Centre's application mix includes:

- Computational Fluid Dynamics: Optimisation of turbines and wings, noise reduction, air conditioning in trains
- Fusion: Plasma in a future fusion reactor (ITER)
- Astrophysics: Origin and evolution of stars and galaxies
- Solid State Physics: Superconductivity, surface properties
- Geophysics: Earth quake scenarios
- Material Science: Semiconductors
- Chemistry: Catalytic reactions
- Medicine and Medical Engineering: Blood flow, aneurysms, air conditioning of operating theatres
- Biophysics: Properties of viruses, genome analysis
- Climate research: Currents in oceans.

Sharing an operational production system between all these stakeholders requires balancing not only the needs of these different research activities, but also the commitments required by the routine, sustained services. Thus supercomputing centre is at the same time both an advanced service provider and a testing ground for the latest (and thus inherently somewhat immature) computational models and other software innovations. This continuous negotiation/consultation process allows supercomputing centres to serve as fulcrums and knowledge exchange channels for broader communities of application developers. As a result, e-Infrastructure service providers reap the benefits in terms of speed and efficiency of the maturing process of the application software – a benefit that is not limited to the supercomputing centres themselves.

In more detail, the key to the fulcrum role lies in the fact that the "grand challenge" supercomputing applications necessitate adopting innovative, experimental approaches both in the modelling software itself as well as the hard- and software infrastructure used to run it. Only by looking at all of these parts together, it is possible to achieve results that exceed both quantitatively and qualitatively the current state of the art. However, typical supercomputing centre is at the same time entrusted with provision of services that have very rigorous, long-term quality of service requirements (e.g. long-term archival of digital cultural heritage artefacts) that cannot be disrupted. The ability to let these two modes of operation – experimental and e-Infrastructure – coexist mean that top-level scientific computing services automatically play a crucial role in the maturing process of scientific software.

In this paper, we present the traditional model of supercomputing application development and two complementary models that leverage this “dual role” of more efficiently to facilitate and speed up the transition of theoretical models into mature computational services. The first of these advanced models is based on workshops arranged on location at the supercomputing centre. The PiCS [5] model represents a longer-term partnership where the computing centre provides a forum where computational modellers and software engineers can develop a common vision for optimal approach to software maturity and requirements for the underlying e-Infrastrucutre. We also present case studies that illustrate these models in action.

IV. Traditional Approach

The traditional approach to scaling up software to top-level computing systems is closely mimicking the general peer-review process. A software developer (i.e. researcher) attempts to prove that his project represents the best return on investment for resources on systems such as supercomputers that are either unique or have demand that outstrips the supply. One of the key arguments in this “formal proof” – in addition to the potential contribution to the research question itself – is the maturity of the software in question, especially in the sense that it can utilise the resources of the top-tier systems efficiently (scalability).

These claims by the researchers are inspected by panels of experts, who are encouraged to discard in their evaluations any knowledge of skills, motivations, previous experience, related initiatives (at the centre or among centre’s clients) that are not brought up in the application itself. After the proposals have been evaluated, the top-ranked ones are given access to the system. The access is based on an account with a quota (that specifies the maximum amount resources to be used etc.) that also gives access to support functions (ticketing system and contact information of the helpdesk) that assist in solving problems related to running software efficiently in the new environment. In case of mature software solutions, this approach is probably necessary to provide equal and fair access to the resources based prioritisation on scientific merit. However, in situations where the research question would require using modelling software that is not yet proven to be on the highest maturity level, the inherent delays of the traditional approach can be problematic. For example, the delay can mean that the team developing the software gets temporarily reassigned to other tasks, making them less effective in tackling the issues as they emerge in the large-scale systems when they eventually receive access.

In the case of PRACE allocations [6], the review process takes place every 6 months for major projects and 4 times a year for smaller scale preparatory access. This means that even if the first application for resources is successful, there will be a considerable (up to a year) delay between the moment when the need for the large-scale systems is identified and the software can be tested in the top-level systems. The cycle for the preparatory access [7] is shorter (evaluations taking place every three months) and thus halves the delay before the basis adaptation of the software to the supercomputing environment can be started. However, some of the issues with the software become apparent only once the problem size and allocated resources are scaled up sufficiently.

Less obvious, but potentially more insidious problem with the traditional approach is the inherent positioning of the parties: the supercomputing service providers are positioned as “gatekeepers”, with the task of keeping all but the “safe”, independently vetted top-applications out of the system. Conversely, the hopeful users-to-be have at least a theoretical incentive to downplay any known issues with the application code. This application-period situation may make some of the researcher less inclined to reach out and access the expertise of the top-level e-Infrastructure experts. In the end, the “time to results” may end up being longer than necessary.

V. First Step: Joint Workshops

To address these shortcomings, Leibniz Supercomputing Centre of the Bavarian Academy of Science and Humanities (LRZ) initiated “Extreme Scaling Worshop” [8] in July 2013. Projects were invited to work together with the supercomputing experts to port their application codes to SuperMUC [9] in a way that utilises efficiently the whole system. The workshop was targeting applications that were already tested in relatively high-end supercomputing environments, thus the application software solutions were relatively mature. However, the goal of the workshop was to scale up to a system 4.5 times the size previously attempted anywhere with the applications in question.

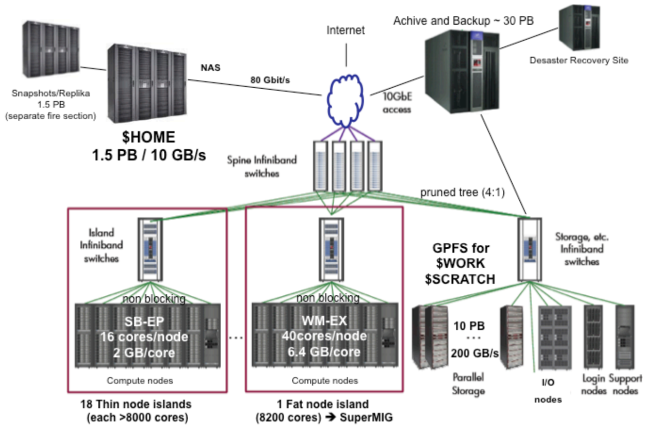


Figure 1. SuperMUC System Architecture

Despite the high level of familiarity both parties had with the environments, applications and tools, the results of the event still highlighted the advantages of the concentrated joint activities: about half of the applications reached the ambitious goal of using the whole SuperMUC system (depicted in the Figure 1 above) efficiently, while almost all of the remaining ones reached the halfway milestone (65 536 cores). In addition, the workshop allowed the application developers and supercomputing centre staff to test very rapidly different optimisation approaches that both made these excellent scaling results possible and resulted in some cases (Gromacs software) in additional 10-15% performance gains through fine-tuning some operational parameters.

The basic outline of the workshop was based on the following model [10]:

- The entire SuperMUC was reserved for the workshop duration of 2.5 days, with 0.5 days reserved for initial testing and 2 days for the execution of the scalability challenges
- LRZ provided automatic tools to automate compilation, submission and validation of application software and its results. This played a key role in making testing different applications in quick succession possible
- Intensive "boot camp" approach was successful in creating in-dept knowledge that the participants could pass on further. As an example, follow-up activities led to an application "performance world record": Seissol [11] software was executed at close to 1 Petaflop/s (i.e. almost 1/3 of the theoretical SuperMUC maximum).

The summary of the increases in the number of cores the software suites can effectively use are presented in the table 1 below. Due to the successful execution of the workshop, it has become a permanent event organised by LRZ on annual basis.

Package	Description	#cores reached	TFLOP/island	TFLOPS (max)
Linpak	Top500 benchmark	128 000	161	2560
Vertex	Plasma Physics	128 000	15	245
GROMACS	Molecular Modelling	64 000	40	110
Seissol	Geophysics	64 000	31	95
waLBerla	Lattice Boltzmann	128 000	5.6	90
LAMPPS	Molecular Modelling	128 000	5.6	90
APES	CFD	64 000	6	47
BQCD	Quantum Physics	128 000	10	27

Table 1. The First Extreme Scaling Workshop results - started level was 32,000 cores (except Linpak).

However, even in light of these excellent results, it is not possible to organise these workshop with much higher than annual frequency. Finding a suitable 2.5 day time period where reserving (almost) the full capacity of the supercomputer is possible usually only few times per year. Similarly, finding dates during which all of the experts at LRZ as well as the key application developers from different projects can all travel to Munich for the duration workshop may prove to be equally difficult. So, despite its clear additional advantages (e.g. cross-pollination between different application domains), the workshop approach cannot – on its own – fulfil the full potential of a supercomputing centre as a catalyst in rapid maturing of scientific software.

VI. PARTNERSHIP INITIATIVE PICS

To address the limitations of workshop-based approach, LRZ has launched the Partnership Initiative π^{CS} (pronounced "pics")[5] to allow more intensive and longer-term collaboration between supercomputing experts and application developers. In the π^{CS} model the supercomputing centre assigns a dedicated contact person for the application scientist, who will – during the course of the long-term relationship – take care of (among other things):

- Arranging suitable execution environments for the software
- Liaising with the software development and quality management support
- Provide training tailored to their specific needs
- Arrange access to exclusive resources, specialised infrastructure or test environments.

Ideally the dedicated contact person has a background in the application science he or she is supporting, which makes it easier to find additional synergies. In the simplest case, this benefits both internal and external communications. For example, common background makes it easy to produce press-releases and other outreach material aimed at presenting the novelty of the achievements – both in terms of IT and the basic research – to the general public. More ambitious modes of collaboration will also be more feasible, e.g. tight collaboration between the computer scientists and computational modellers will make joint publications of new joint research activities considerably easier to launch.

This approach essentially brings in one of the key approaches of mature service management (customer relationship management) as a complement to – or in some cases a replacement of – the somewhat impersonal traditional approach described in chapter IV. While this change doesn't perhaps seem that remarkable in itself, it implicitly introduces new metric to the computing centre's management practices: partner satisfaction and/or partner research results. Eventual monitoring and tracking of this additional metric will in turn make it much easier to present the intuitive understanding of the added value of the specific services provided. If adopted as a formally recognised metric, the data can be presented to funding agencies and other stakeholders in the centre's governance to provide additional input for the strategic decision-making and evaluation of centre's efficiency.

However, in the case of the leading-edge supercomputing applications, we can also assume that the customer satisfaction is also tied to the improvements in the scalability of the software. Against this backdrop, we can postulate that the additional improvements to the Extreme scaling workshop "graduate software suites" illustrate the additional improvements made possible by the partnership model. For example, the Seissol application that was scaled from 32k to 64 cores was further developed based on the partnership model. This made it possible to reach the performance of over 1400 TFLOPS using over 145 000 cores. This corresponds to 44.5% of the theoretical peak performance, which is extremely rare achievement for an actual application code.

VII. PiCS in Action: DRIHM Project Experience

The underlying theories and approaches behind the π^{CS} initiative were tested and refined in the final stages of the DRIHM (Distributed Research Infrastructure for Hydro-Meteorology) project [12]. The goal of the project was to create an open, fully integrated workflow platform for predicting, managing and mitigating the risks related to extreme weather phenomena. Reaching this goal required integrating numerous different components into multi-model chains that could be executed automatically to analyse e.g. flood risks (both statistical and based on specific event scenarios) and other hydrometeorological research (HMR) challenges.

DRIHM project ran from 1st September 2011 to 28th February 2015, with consortium involving partners representing major computing centres, hydrometeorological model developers and organisations with operational responsibilities (most notably Republic Hydrometeorological Service of Serbia). LRZ was involved through MNM (Munich Network Management) Team [13] that links LRZ and several academic institutions in the Munich area.

HMR has been identified as one of the key domains where improvements in the speed and accuracy of modelling would have a profound socioeconomic impact. In the 2013 paper [14] average global annual flood losses are estimated as $6 billion in 2005, estimated to increases somewhere between $52 billion and one trillion dollars due to socio-economic and climate changes. Thus the DRIHM project represented an application domain where socioeconomic factors clearly supported using the highest level of computing resources, even in the case where all the software components were not necessarily at the highest possible maturity level.

The starting point of the project was quite challenging: it was not only necessary to develop the software/metadata/procedural framework to link the model components together during the project lifetime, but the models themselves were in some cases used only independently (i.e. not as part of a multi-model system) in the desktop computer or small-scale computing cluster settings. The top part of the Figure 2 presents the starting point of the project, where each of the models tended to have specific demands for the execution environment and produced output files that were not compatible with each other.

The requirement gathering and design work was conducted with an approach that resembled a combination of workshop and π^{CS} models. The foundations were built in a series of project workshops or working meetings in order to ensure that the common design could accommodate all of the models (both currently identified and ones anticipated as candidates for future integration). Once these foundations were laid out, the work progressed with the participating computing centres and computer science competence centres providing named experts as contact points for the model developers and operational organisations.

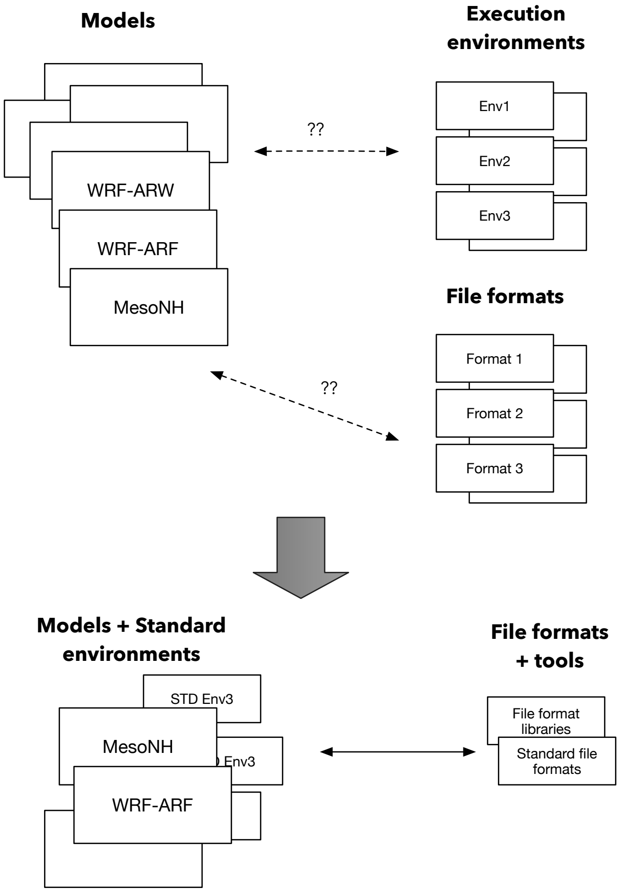


Figure 2. Modelling software framework development during the DRIHM project

In collaboration with these experts, the project developed approaches (so called M.A.P approach [15]) that made it possible to create and maintain consistently the different execution environments required by the component models as well as to perform necessary data conversion operations in a way that was verifiable in terms of syntactic and semantic compatibility (the lower part of the Figure 2). The project results validated the key assumptions behind the π^{CS} model:

- Partnering the experts together clearly reduced (in some cases eliminated) the gap between computational researchers and IT experts
- The technology and knowledge transfer was extremely efficient, allowing the project reach its goals. The automatic execution (in a matter of minutes or few hours) of a model chain that previously took weeks or even months to perform. This made it possible to consider approaches that were previously strictly limited to post-mortem analysis in operational role (early warning, steering the response etc).
- The intense interdisciplinary collaboration made a very successful "summer school" possible, which in turn triggered numerous follow-up actions.

The pairing approach also allowed a very flexible approach in terms of working meetings – during the last year several meetings scheduled on demand to address challenging issues. The success of this approach is also evident in the ability to scale some of the codes from few hundred cores to a level where running them on SuperMUC was justified.

VIII. FUTURE DIRECTIONS

The future approach will be based on building on the successful application of the model in the HMR domain and extending the approach to the broader "Environmental Computing" [16] area. There are a large number of activities where problems studied are inherently interdisciplinary in nature and thus need an efficient and coherent multi-modelling approach. Expanding the number of disciplines and sub disciplines will increase the demands for the systematic collection and curation of metadata related to the model components. However, meeting this challenge is necessary in order to respond to new challenges in disaster risk reduction and other crucial activities of societies that rely on accurate modelling of natural phenomena. These activities are becoming more and more important and visible as the risk landscape is changing dramatically (due to climate change, for example). The intergovernmental response to these challenges is leading to new policies – such as Sendai declaration [17] – that give civil defence actors new mandates as well as new responsibilities.

Partnering with organisations that have sufficient visibility and credibility to drive common approach forms another strategic component in the application of the π^{CS} model. LRZ has an ongoing collaboration with United Nations Office for Disaster Risk Reduction (UNISDR) [18] in the computationally intensive tasks related to the next version of the Global Assessment Report [19]. The role of UNISDR is to collect input from different modelling activities and bring them together to produce an overview document that can be used to assess the efficiency of the disaster risk reduction policies as they have been implemented by the UN member states. Through UNISDR collaboration LRZ gains access to a very large and diverse network of experts developing models for all the different environmental disaster risk scenarios considered in the assessment report. This makes is considerably easier to identify common requirements and approaches used by the model developer community. This contact network can also be used to promote successful approaches (such as using LRZ Cloud services [20]) to a larger group of users.

The role Cloud and other virtualisation solutions will also grow more prominent, as the rapid development cycle that characterises the π^{CS} greatly benefits from the ability to tune the virtual execution environment to be as close to the original development environment as possible. This should also make it easier to adapt to the eventual situation where the "native" development environment of the environmental computing models will be the Cloud instead of the desktop. We would like to emphasise that virtualisation of the resources and HPC services that Cloud represents does not remove the challenges addressed by the two approaches presented. In fact, the "pay as you go" approach and the lack of "steps" in the capacity/price ratio where the incremental costs suddenly becomes very high (e.g. due to the need to rewire the hosting space or additional cooling requirements) may well lead to a situation where it is tempting to increase hardware resources instead of optimising the code. This will be possible up to a point, but the scalability and reliability issues will be considerably more challenging when they are (eventually) encountered. Furthermore, dealing with these issues may be more difficult due to more transient relationship with the support organisations (e.g. if the software has been run on different Cloud environments).

IX. CONCLUSIONS

Software is a crucial component in both modern scientific discovery and in providing reliable, well-designed and coherently managed IT services to support research and other human endeavours. Somewhat paradoxically, the approaches to developing these crucial components differs considerably depending on whether their role is to directly support scientific discovery or if they are aimed at providing infrastructure-level solutions for broader audience. Nevertheless, almost all of the successful infrastructure solutions have started their life – at least conceptually – as research projects, and made the (sometimes painful) transition to maturity through a fairly ad hoc process.

Supercomputing centres have a great potential to both facilitate and steer this maturing process. As the nature of their day-to-day operations incorporates aspects of both leading edge research and mature service provision, they are probably uniquely placed to act as bridge builders between the computational modelling in the basic research and software engineering and service management in more formalised services. Unfortunately the traditional approach for granting access to hardware resources – and especially to support services (such as scalability consulting and technical support) – misses quite a lot of this potential.

Based on the experiences presented in this paper, we expect the traditional approach to be complemented more and more often with partnership-oriented approaches, both as intensive, high-profile workshops and longer term partnerships. These approaches are most likely to become more and more critical as the traditional IT support - researcher relationships are challenged by the Cloud-based approach. Somewhat paradoxically we estimate that the model where the resource constraints are more efficiently hidden from the end users, the training, consulting and other services that allow researchers to use these new resources more efficiently should be offered proactively earlier in the product lifecycle than in the traditional approach to provisioning of IT services for researchers.

REFERENCES

[1] GitHub service homepage, https://github.com/ (accessed 4th June 2015)

[2] SoftwareX Journal homepage, http://www.journals.elsevier.com/softwarex/ (accessed 4th June 2015)

[3] e-Infrastrucure Reflection Group White Papers, 2003 – 2013, http://e-irg.eu/white-papers (accessed 4th June 2015)

[4] e-Infrastructure Reflection Group, “Best Practices for the use of e-Infrastructures by large-scale research infrastructures”, http://e-irg.eu/documents/10920/277005/Best+Practices+for+the+use+of+e-Infrastructures+by+large-scale+research+infrastructures.pdf, p17 (accessed 4th June 2015)

[5] A. Frank, F. Jamitzky, H. Satzgera, and D. Kranzlmüller, “In Need of Partnerships - An Essay about the Collaboration between Computational Sciences and IT Services”, Procedia Computer Science, Vol 29, pp 1816 – 1824, doi:10.1016/j.procs.2014.05.166

[6] PRACE Research Infrastructure, “Resource Allocation”, http://www.prace-ri.eu/resource-allocation/ (accessed 4th June 2015)

[7] PRACE Research Infrastructure, “PRACE Preparatory Access”, http://www.prace-ri.eu/prace-preparatory-access/ (accessed 4th June 2015)

[8] M. Allalen et al, “Extreme Scaling Workshop at the LRZ”, Advances in Parallel programmin, Vol 25, pp 691 – 697, DOI 10.3233/978-1-61499-381-0-691

[9] Leibniz Supercomputing Centre, “SuperMUC Petascale System”, http://www.lrz.de/services/compute/supermuc/ (accessed 4th June 2015)

[10] Presentation: D. Kranzlmüller, “Extreme Scaling on SuperMUC”, slide 25 onward, available from: http://www.mnm-team.org/_vortraege/Kranzlmueller_20131216_Barcelona-Extreme_Scaling_on_SuperMUC.pdf (accessed 4th June 2015)

[11] Seissol working group, “High Resolution Simulation of Seismic Wave Propagation in Realistic Media with Complex Geometry”, http://seissol.geophysik.uni-muenchen.de/ (accessed 4th June 2015)

[12] DRIHM project, homepage, http://www.drihm.eu/ (accessed 4th June 2015)

[13] MNM Team, homepage, http://www.mnm-team.org/ (accessed 4th June 2015)

[14] S. Hallegatte, C. Green, R. J. Nicholls, and J. Corfee-Morlot, “Future flood losses in major coastal cities”, Nature Climate Change, Vol 3, pp 802 – 806, doi:10.1038/nclimate1979

[15] DRIHM project, “Tutorial on adapting your model to DRIHM”, Available from: http://www.drihm.eu/images/SupportCentre2014/DRIHM_MAP.pdf (accessed 4th June 2015)

[16] LMU & LRZ, “Environmental computing homepage”, http://www.envcomp.eu/ (accessed 4th June 2015)

[17] United Nations General Assembly, “Sendai Framework for Disaster Risk Reduction 2015 – 2030”, A/CONF.224/L.2, Available from: http://www.wcdrr.org/uploads/Sendai_Framework_for_Disaster_Risk_Reduction_2015-2030.pdf (accessed 4th June 2015)

[18] The United Nations Office for Disaster Risk Reduction UNISDR, homepage, http://www.unisdr.org/ (accessed 4th June 2015)

[19] UNISDR, “Global Assessment Report”, Available from: http://www.unisdr.org/we/inform/gar (accessed 4th June 2015)

[20] LRZ, “The LRZ Compute Cloud”, Available from: http://www.lrz.de/services/compute/cloud_en/ (accessed 30th August 2015)

Health Professional Students' Acceptance of Mobile Information Communication Technologies for Learning -

a Study Using the Unified Theory of Acceptance and Use of Technology (UTAUT)

Lili Liu, Shaniff Esmail, Elizabeth Taylor
Department of Occupational Therapy
University of Alberta
Edmonton, Canada
e-mail: lili.liu@ualberta.ca, shesmail@ualberta.ca,
liz.taylor@ualberta.ca

Adriana M. Ríos Rincón, Antonio Miguel Cruz
Department of Occupational Therapy
University of Alberta
Edmonton, Canada
School of Medicine and Health Sciences
Universidad del Rosario
Bogotá, Colombia
e-mail: aros@ualberta.ca/adriana.rios@urosario.edu.co
miguelcr@ualberta.ca/antonio.miguel@urosario.edu.co

***Abstract*— The aim of this study was to examine what factors affect the acceptance behavior of mobile information communication technologies for learning by students in a MSc Occupational Therapy program in a Canadian university. The study addresses mobile and distance education, specifically, the function of mobile learning in higher education. A self-administrated paper-based survey was created by adapting scales used in previous research based on the Unified Theory of Acceptance and Use of Technology (UTAUT). Our research model was tested using the Partial Least Squares (PLS) technique. *Social Influence* was the strongest salient construct for *behavioral intention* to use mobile information communication technologies for learning, followed by *performance expectancy* of mobile information and communication technologies. *Effort expectancy* was not a salient construct for behavioral intention to use these technologies.**

Keywords-m-learning; UTAUT; health sciences; education.

I. INTRODUCTION

Occupational therapists are regulated health professionals that work in a variety of public and private settings to address physical and psychiatric (mental health) issues to help people or populations maintain or return to their regular activities including work, leisure and self-care. In 2011, there were approximately 1,532 occupational therapists practicing in Alberta and 13,501 practicing in Canada [1]. The occupational therapy program is about two years (26 months) and located in Edmonton. In 2012, the Department of Occupational Therapy created a satellite program in Calgary which is 300 Km away south of Edmonton. At any one time, there are two cohorts (year one and year two). Each cohort has 98 students in Edmonton and 22 students in Calgary. The program uses distributed learning approach where students at both sites receive instruction at the same time and follow the same curriculum through the use of high definition videoconferencing and other information and communication technologies.

Information and communication technologies have the potential to enhance education and professional networks between healthcare professional including occupational therapists [2]. Our aim as educators is to prepare future generations of health professionals to perform their jobs in the community and in clients' home environments. In essence, we wish to apply information and communication technologies to create meaningful, context-specific, community-based learning for our students who will work in teams or communities of practice. Information and communication technologies have been used for learning in a variety of disciplines. E-learning is based on the use of wired and wireless Internet. Mobile learning (or m-learning) is a part of e-learning. Mobile learning is defined as "any sort of learning that happens when the learner is not at a fixed predetermined location, or learning that happens when the learner takes advantage of the learning opportunities offered by mobile technologies" [4, p. 6]. Mobile learning refers to learning activities facilitated by the use of mobile information and communication technologies, such as cell phones, smart phones, palmtops, tablet personal computers, personal digital assistants and portable multimedia players [3]. Mobile information and communication technologies have the potential to provide educational opportunities for students in higher education because they can facilitate students' access to information and interaction with instructors, peers and colleagues regardless the place where they are located [5]. Mobile information and communication technologies are expected to support the learning experience in several ways. They can: support, guide, and extend the students' thinking process within and out of the classroom; enhance learner creativity, exploration and problem solving; facilitate the process for students to express their opinions; and enable learning with students' preferred approach and

ISBN: 978-1-61208-432-9

speed of communication, making learning more autonomous and self-reliant [6].

University students value the portability and immediacy of smart phones and tablets for obtaining and sharing information with peers [7]. Mobile information and communication technologies have been accepted by students in university lectures and are perceived as useful and easy to use. Additionally, students' attention and motivation are higher with metacognitive supports provided via mobile technologies during class [8]. Research has revealed that the use of technologies for learning activities depends on students' perception towards technologies [9]. However, little research has investigated the factors that determine students' acceptance of mobile information and communication technologies for learning [3]. Much less attention has been given to the study of factors associated with adoption behavior of mobile information and communication technologies by occupational therapy students.

Theories from the social sciences have explained how and why people adopt technologies, calling this construct as the behavioral intention to use technology [10]. The Unified Theory of Acceptance and Use of Technology (UTAUT) [11] integrates previous models with the behavioral intention perspectives and use of technologies. The UTAUT model has six constructs: 1) Performance expectancy (PE) defined as the degree to which a person believes that using the technology will help him or her to attain gains in job performance. 2) Effort expectancy (EE), the degree of ease associated with the use of the technology. 3) Social influence (SI), the degree to which a person perceives that important others believe he or she should use the technology under study. 4) Facilitating conditions (FC), the degree to which a person believes that an organizational and technical infrastructure exists to support use of the technology. 5) Behavioural intention (BI), the intention to do some Behavior and 6) Use, the overt behavior [10].

According to the UTAUT model, four constructs play a role as direct predictors of behavioral intention to use the technology under study and two have a direct influence on the use. Performance expectancy (PE), effort expectancy (EE), and social influence (SI) are direct determinants on *behavioural intention* (BI); and facilitating conditions (FC) and Behavioural Intention (BI) to use the technology are the two determinants that have a direct impact on *use* of the technologies. Based on this theoretical framework, our study objectives were to develop a path model (path analysis) of mobile information and communication technologies acceptance by university students and to analyze the relationship of the UTAUT constructs, i.e., how performance Expectancy, Effort Expectancy, Social Influence determine students' Behavioural Intention to use mobile information and communication technologies for learning. In addition, some descriptive statistics related to m-learning use were also used to explain the Behavioural Intention. We think that our results might help program administrators and instructors implement strategies for m-learning. Thus, the aim of this study was to answer the following research question:

What factors affect the acceptance behavior of mobile information communication technologies for learning by students in a MSc in occupational therapy program in a University in Canada?

This paper is organized as follows: related works, the theoretical framework and the research objectives and question are presented in the first section. Materials and methods are presented in the second section. Results, discussion and conclusions are presented in the third, fourth and fifth section, respectively.

II. MATERIALS AND METHODS

In this study, we used a cross-sectional exploratory approach using a self-administrated paper-based survey. This study received approval from the University of Alberta Health Research Ethics Board. The target population consisted of all students at the occupational therapy masters program (MScOT). We tested the model using the Partial Least Squares (PLS) technique. We adopted guidelines for sample size where a minimum of 10 subjects should be surveyed per total number of dependent variable with the largest number of independent variables influencing it [12]. Therefore, the minimum sample size required for this study was 40 subjects.

We created a survey questionnaire designed to measure the constructs and relationships contained in our research model. The 36-items were grouped into three sections. In the first section (section A1, items from 1-5), we asked for participant demographics and background or previous degree. In the second section (section A2, item 6 and 7) we inquired about the students' experience using mobile information and communication technologies including mobile technology (e.g., smart phone, tablet, digital camera), and mobile applications and software (e.g., paid short message service, Skype, wiki), as well as the average daily use over the last week (in hours/day) of these technologies during personal and study time. In the third section (section B, items from 8-35), we created specific questions (items 8-35) by adapting scales and items already validated and with high levels of internal consistency in previous research for each construct of the model (Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions and Behavioural Intention). Part B also included four items for assessing attitude towards and anxiety generated by mobile information and communication technologies [3] [11].

Under the logic of PLS modeling, the model is formed by one or more blocks, which is a structure formed by a latent variable (each UTAT construct in this study) with its manifest variables (MV) (each questionnaire item per construct). The LVs can be exogenous, endogenous, or both. A latent variable is exogenous when it is not predicted by any other latent variable. A latent variable is endogenous when it is predicted in by one or more latent variables. The structural model is composed of the relationships or paths amongst exogenous and endogenous latent variables. In this research we had one outcome (endogenous) latent variable (i.e., Behavioural Intention); and four independent latent variables (Performance Expectancy (questionnaire items 8-11), Effort Expectancy (items 12-15), and Social Influence (items 16-19) were considered as direct latent variable (independent or exogenous variables) of Behavioural Intention. We measured Facilitating Conditions (items 20-23), however, we did not use it in the path model because Facilitating Conditions is considered to be direct latent variable of use in the UTAUT model and we did not measure use behavior in this study. We measured attitude and anxiety toward using technologies, however, we did not include them in the path analysis in our study because it has been reported that these constructs are significant only when performance and effort expectancies are not included in the model [11].

Confounding variables were operationalized according to [13] methods as follows: Dichotomous variables were coded as "0" or "1" (e.g., student's gender, and student's year in the MScOT program). All items of section B of the questionnaire (items from 8-35) related to each dependent and independent latent variables were scored on a 7-point Likert scale, ranging from "strongly disagree (1)" to "strongly agree (7)" [14]. As we had four questions per construct, the numbers of items per construct exceeded the minimum three items required for proper calculation of measurement errors [15].

Before sending out the surveys to participants we conducted a pilot study with six students (graduate and undergraduate) selected by convenience. We made minor changes to clarify some of the questionnaire items. We held a meeting with all potential study participants (230 MScOT students) where researchers distributed packages, including an information letter, informed consent form and the survey questionnaire to students. Students who agreed to participate filled out the survey at that moment. We used codes for the surveys instead of students' names or ID in order to ensure anonymity and confidentiality. Students who were unable to attend the meeting were invited to sign the consent form and complete the questionnaire online. A list with the match between students' identifiers and survey codes were stored in a locked file cabinet.

Before the statistical analysis, a random sample of 20% of the entered data was compared to check coding accuracy. We used descriptive statistics to summarize demographic data. We conducted correlation analysis (Pearson or Spearman Rho as appropriate) to determine whether survey responses for Performance Expectancy, Effort Expectancy, Social Influence and, Behavioural Intention to use, were correlated with students' age, gender, city of the program, year in the program, attitude and anxiety toward technologies, and average daily hours of mobile information and communication technologies use for personal purposes, and education purposes. Missing data was handled in the following manner: (1) missing data of continuous variables such as students' clinical experience and age were replaced by the average values of these variables; and (2) for categorical and ordinal variables such as the discipline or type of the therapists, and their educational level, missing data were replaced by the medians of these variables.

We tested the multivariate research model using the PLS technique [16]. The PLS measurement model evaluation was conducted by means of: (1) reliability measurement for each construct (Cronbach's alpha); (2) convergent validity measurement of each set of items with respect to their associated construct will be assessed by examining the factor loadings of the items on the model's constructs; and (3) discriminant validity was analyzed by using Average Variance Extract (AVE) indicator. PLS structural model was evaluated by means of (β) paths coefficients, the explained variance (R^2) and the effect size (f^2) for each path segment of the model. Also the Bootstrapping re-sampling method was employed to verify the statistical significance of (β) paths coefficients of the PLS model. The alpha level of significance was set at $p \leqslant 0.05$. IBM SPSS® V 22.0 and SmartPLS V 2.0 M3statistics package were used to generate descriptive, univariate and bivariate statistics, and PLS path modeling respectively.

III. RESULTS

Regarding our sample size, considering that the potential subjects to be surveyed were 230 students, and that we retrieved 213 surveys, we achieved a 93% response rate and a statistical power of 100%. The only large effect size was the one for the path Social Influence→ Behavioural Intention: (f^2=0.771) (see Table II). We achieved 99.5% of accuracy in data entering. Missing data was low (e.g., we had missing information in the students' year (0.9% of respondents) and previous experience with of some information and communication technologies (between 0.9% and 2.2% of respondents). In part B of the survey, we had 3.8% missing data. After using missing data procedures we found negligible changes in variables.

TABLE I. STUDENTS PREVIOUS EXPERIENCE WITH TECHNOLOGIES

ICTs used	n (%)	AHP[S.D.]	AHE[S.D.]
Mobile ICTs	204 (96.2)	9.00[9.00]	6.08[4.12]
Smart phone	200 (94.3)	3.3 [3]	1.1 [1.8]
Mobile phone	6 (2.8)	2.5 [2.1]	5.5 [6.4]
Laptop computer	203 (95.8)	2.5 [2.3]	4.3 [2.3]
Tablet	56 (26.4)	1.2 [0.9]	1.7 [1.9]
GPS navigation device	49 (23.1)	0.7 [1.1]	0.1 [.31]
Audio/Video recording	12 (5.7)	0.6 [0.5]	0.0 [0.1]
Digital camera	33 (15.6)	0.7 [1.3]	0.5 [1.6]
Other device	5 (2.4)	1.3 [1.5]	0.7 [1.2]
Paid short message service	143 (67.5)	2.2 [3.9]	0.5 [2.6]
Free mobile messaging app	84 (39.6)	1.6 [3.0]	0.1 [0.8]
Goniometer App	3 (1.4)	0.0 0	0.00
Skype	64 (30.2)	1.5 [2.7]	0.0 [0.1]
FaceTime	63 (29.7)	0.8 [0.7]	0.0 [0.0]
Blogs	20 (9.4)	1.2 [1.2]	0.8 [1.3]
Wiki	35 (16.5)	0.7 [0.9]	1.0 [1.2]
Other app	5 (2.4)	1.7 [2.5]	1.3 [1.1]

S.D: Standard deviation, **Sample size: 212**
AHP: Average hours ICTs daily use for personal purposes
AHE: Average hours ICTs daily use for education

Overall, participants had an average age of 24.81years (SD 9.93), were mainly female (90.6%), and located in Edmonton (82.5%). Table I shows the previous experience of students with mobile devices and applications as well as students' average daily hours of mobile information and communication technologies (devices and applications) for personal use in the last week. Overall: (1) almost all students had previous experience using mobile devices and applications; (2) the mobile devices most used by students were smart phones (94.3%) and laptop computers (95.8%); (3) the mobile applications most used by students were paid short message service (SMS) (67.5%), free mobile messaging app (39.6%) followed by Skype and FaceTime (30.2% and 29.7% respectively). Smart phones had the highest average weekly hours of personal use (3.3 hours), Mobile phones and laptop computers had the highest average of daily hours used for education (5.5 hours, and 4.3 hours respectively).

We used the UTAUT constructs to examine the overall perceptions of students about mobile information and communication technologies for learning: (1) students thought that mobile information and communication technologies for learning will help them to increase their academic performance and learning (Performance expectancy: 78.2% Agree (Agree-strongly agree), Mode 5.00; Mean 5.32 SD 1.20); (2) students perceived that mobile information and communication technologies for learning are easy to use or not complicated to use (effort expectancy: Agree (Agree-strongly agree: 76.7%), Mode: 5, Mean 5.30 SD 1.08); (3) students tended to be either neutral or agree with the perception that the intention to use mobile information and communication technologies for learning is influenced positively by the opinions and perceptions of peers or instructors. (Social Influence: agree (Strongly-Agree: 35.9%), Neither agree or disagree: 37.5%), Mode: 4, Mean 4.14 SD 1.20); (4); students agreed that in the academic program under study, most of the conditions such as opportunities, resources, technical support and knowledge, as well as that the mobile information and communication technologies are compatible with their educational goals (facilitating conditions Agree (Agree-Strongly: 74.6%), Mode: 5, Mean 5.07 SD 1.18); (5) there was a strong trend in Behavioural Intention to use mobile information and communication technologies for education by students (Behavioural Intention: Agree (Agree-Strongly: 72.2%), Mode: 5, Mean 5.14 SD 1.26); (6) in the same way students´ attitude towards using mobile information and communication technologies for learning is positive (Attitude: Agree (Agree-Strongly: 65.2%), Mode: 5, Mean 4.88 SD 1.14); and (6) students disagree that mobile information and communication technologies generate anxiety in terms of apprehension, intimidation, hesitation or stress (Disagree (strongly-Disagree: 53.6%), Mode: 3; Mean: 3.53, SD: 1.42).

Although the UTAUT model includes gender, age, experience with the technology and voluntary use as possible moderators in the relationship between the four main constructs and the Behavioural Intention or use of technologies [11], we did not include age because age was homogeneous in our sample. Neither did we include gender because in a bivariate analysis the correlation was not significant (Spearman Rho: 0.049, p=0.24). As a measure of experience in the use of mobile information and communication technologies, we included as confounder variable in the PLS multivariate analysis the average hours of use of mobile information and communication technologies for education whose correlation with behavioral intention was found to be significant (Spearman Rho: 0.398, p=0.02).

The results of the structural model estimate are shown in Table II. We ran the PLS structural model using the bootstrap procedure with 500, 1000, 2000, and 5000 times of resampling and the magnitude and significance of the structural paths were consistent. The multivariate model in PLS structural model showed that: (1) that there is statistically significant and positive correlation between Performance Expectancy (PE) and Behavioural Intention (BI) to use mobile information and communication technologies for learning (PE→BI=+0.237, p<0.000); (2) there is no statistical evidence to support the assertion that Effort Expectancy (EE) has a positive influence on Behavioural Intention (BI) to use mobile information communication technologies for learning (EE→BI=+0.090, p<0.119); and (3) there is a strong statistically significant and positive correlation between Social Influence (SI) and Behavioural Intention (BI) to use mobile information communication technologies for learning (SI→BI =+613 p<0.000). Thus, Performance Expectancy and Social Influence constructs matter in Behavioural Intention to use mobile information communication technologies for learning by students, and whereas Effort Expectancy construct did not.

TABLE II. STRUCTURAL MODEL. (PERFORMANCE EXPECTANCY (PE), EFFORT EXPECTANCY (EE), SOCIAL INFLUENCE (SI), BEHAVIOURAL INTENTION (BI), AVERAGE DAILY HOURS ICTS USE FOR EDUCATION (AHE))

Path	Path Coefficient β	t-value	f^2	Q^2	R^2
PE→ BI	0.237	3.369**	0.073	0.475	0.521
EE→ BI	0.090	1.561	0.011		
SI→ BI	0.613	6.361**	0.771		
AHE→ BI	0.108	1.809	0.024		

Endnotes

* p<0.05; **p<0.01; **f^2**: effect size

$$f^2 = \frac{R^2_{\text{included}} - R^2_{\text{excluded}}}{1 - R^2_{\text{excluded}}};$$

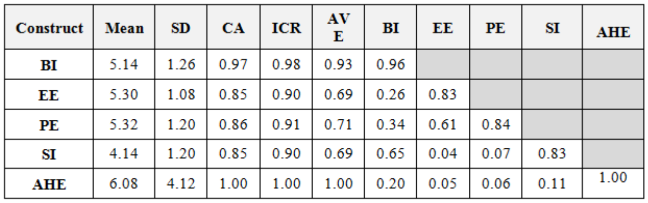
Q^2: Stone Geisser indicator ; **GoF:** Goodness of fit

$$GoF = \sqrt{\overline{Communality} * \overline{R^2}}$$

Regarding the model validity and reliability, all item loadings were statistically significant at the 0.001 level and all item loadings were greater than 0.70, indicating good convergent validity at the indicator level. All internal composite reliability (ICR) values were greater than 0.70, indicating acceptable reliability. The square root of each average variance extracted (AVE) (shown on the diagonal in

Table III) is greater than the related inter-construct correlations in the construct correlation matrix, indicating adequate discriminant validity for all of the reflective constructs.

TABLE III. Construct correlations (Performance expectancy (PE), Effort Expectancy (EE), Social Influence (SI), Behavioural intention (BI), Average daily hours ICTs use for education (AHE)) SD= Standard deviation; CA=Cronbach's Alpha.



Construct	Mean	SD	CA	ICR	AVE	BI	EE	PE	SI	AHE
BI	5.14	1.26	0.97	0.98	0.93	0.96				
EE	5.30	1.08	0.85	0.90	0.69	0.26	0.83			
PE	5.32	1.20	0.86	0.91	0.71	0.34	0.61	0.84		
SI	4.14	1.20	0.85	0.90	0.69	0.65	0.04	0.07	0.83	
AHE	6.08	4.12	1.00	1.00	1.00	0.20	0.05	0.06	0.11	1.00

The explained variance of the model (R2) was 0.521 for Behavioural Intention to use mobile information communication technologies for learning which do appear to be strong according to [17] criteria. The Stone–Geisser's Q2 value for Behavioural Intention to use mobile information communication technologies for learning construct was 0.475, indicating good predictive relevance of our model (Q>0 indicates good predictive relevance).

IV. DISCUSSION

The aim of this study was to examine what factors affect the acceptance of mobile information and communication technologies for learning by students in a MSc program in occupational therapy at a University in Canada. We found statistical support to assert that performance expectancy (PE) and social influence (SI) affect the behavioural intention (BI) to use mobile information and communication technologies for learning. In our study, effort expectancy was not a determinant of Behavioural Intention to use mobile information and communication technologies. Previous research in acceptance of mobile information and communication technologies for learning are mixed. On one hand, a study using the UTAUT model with university students found that Performance Expectancy, Social Influence and Effort Expectancy were determinants of Behavioural Intention to use the free mobile messaging app for learning purposes [18]. However, another study using the Technology Acceptance Model (TAM) found that neither perceived usefulness (Performance Expectancy in the UTAUT) nor perceived ease of use (Effort Expectancy in the UTAUT) had an effect on the Behavioural Intention of students who were taking e-learning courses [3]. In our case, our results that Performance Expectancy and Social Influence determine Behavioural Intention are aligned with the UTAUT model. The fact that Effort Expectancy did not determine the Behavioural Intention in our model can be explained by the fact that the constructs Effort Expectancy and Performance Expectancy were significantly correlated (Spearman Rho: 0.579, p=0.000), thus, these two constructs showed collinearity. In order to control this collinearity, we calculated a structural model in which Performance Expectancy was eliminated. This resulted in Effort Expectancy to become in a statistically significantly predictor of Behavioural Intention; however, without Performance Expectancy the model prediction was reduced in 8% (R2=0.487) which is not convenient. Thus, we decided to keep in our model both Performance Expectancy and Effort Expectancy with a R2 of 0.521.

Students in our sample thought that mobile information and communication technologies have the potential to help them to increase their academic performance and learning. This result is consistent with previous research where university students in Canada and the USA tend to believe that mobile information and communication technologies such as cell phones, smart phones, and tablets are important to their academic success and use their devices for academic activities [5]. Students in our study also believed that mobile information and communication technologies for learning were easy to use, or not complicated to use (EE). This positive perception can be explained by the previous experience with the use of mobile information and communication technologies (mainly smart or mobile phones, laptops, SMS, WhatsApp and Skype) by students as we found a statistically significant correlation between effort expectancy (EE) and the average hours students used information and communication technologies for personal use (Spearman Rho: 0.213, p=0.005). In the same way, students demonstrated a positive attitude towards using mobile information and communication technologies for education and felt that they have the conditions (e.g., resources, opportunities and technical support) for using mobile information and communication technologies for learning. These results are encouraging for academic purposes because we can assume that students perceived that they have the basic skills and conditions for using mobile information and communication technologies. This can ease the implementation of learning strategies using information and communication technologies. Regarding location, we found a statistically significant negative correlation between city and Behavioural Intention, i.e., students in Calgary had higher behavioural intention (BI) to use the mobile information and communication technologies for learning. This result is not surprising because Calgary is a satellite program, Calgary students have the need to do more remote interactions with their instructors in Edmonton than those students living in Edmonton.

On the other hand, we found that students in our study used mobile phones and laptop computers for education more than four hours per day. Other technologies such as tablets, smart phones, free mobile messaging apps, Skype, blogs and wikis were used less than 2 hours per day for education activities despite the literature reporting that these types of mobile information and communication technologies increases collaborative learning, leadership [19], immediacy for obtaining and sharing information with peers [7], and enhancing attention and motivation during lectures [8]. Therefore, our results invite reflections about the need for universities to increase the academic activities

ISBN: 978-1-61208-432-9

where students have opportunities to benefit from the use of information and communication technologies.

We propose further research to examine the additional mobile information and communication technologies to investigate how these strategies can facilitate small group learning approaches and interactions across distances. These strategies include the use of mobile technologies and apps for development of competencies in interviewing simulated clients with mental health conditions, and physical assessments of activities of daily living.

V. CONCLUSION

This study shows that performance expectancy (PE) and social influence (SI) affect the behavioural intention (BI) to use mobile information and communication technologies for learning by students in a MSc in occupational therapy program at a University in Canada. Our structural model achieved a strongly explained variance of Behavioural Intention, good convergent validity and acceptable reliability. In general, students had a positive attitude towards the use of mobile information and communication technologies for learning. However, currently they are using few mobile information and communication technologies devices and applications for academic purposes. Our results support the development of strategies to increase the use of mobile information and communication technologies for teaching and learning with university students in a health profession program.

ACKNOWLEDGMENT

This study was funded by a University of Alberta Teaching and Learning Enhancement Fund (TLEF) 2013-2015. The authors thank Miki Kobayakawa and Pranshu Arora for their assistance with data entry.

REFERENCES

[1] Canadian Institute for Health Information, "Occupational Therapists in Canada, 2011 Occupational Therapist Provincial Reports [Online]. Available: https://secure.cihi.ca/estore/productFamily.htm?locale=en&pf = [Accessed 20 May 2015].

[2] T. Hoffmann, L. Desha, and K. Verrall, "Evaluating an online occupational therapy community of practice and its role in supporting occupational therapy practice," Australian Occupational Therapy Journal, vol. 58, p. 337–345, 2011.

[3] S. Park, M.-W. Nam, and S.-B. Cha, "University students' behavioral intention to use mobile learning: Evaluating the technology acceptance model," British Journal of Educational Technology, vol. 43, no. 4, p. 592–605, 2012.

[4] C. O'Malley, G. Vavoula, J. Glew, J. Taylor, M. Sharples, and P. Lefrere, "Guidelines for learning/teaching/tutoring in a mobile environment," MOBIlearn/UoN,UoB,OU/D4.1/1.0, n.p., 2003.

[5] J. Gikas and M. Grant, "Mobile computing devices in higher education: Student perspectives on learning with cellphones, smartphones & social media," Internet and Higher Education, vol. 19, p. 18–26, 2013.

[6] J. Rikala and M. Kankaanranta, "The Nature Tour Mobile Learning Application Implementing the Mobile Application in Finnish Early Childhood Education Settings," in 6th International Conference on Computer Supported Education CSEDU 2014, Barcelona, Spain, April 2014, vol. 3, pp. 171-178,.

[7] P. Rebaque-Rivas, E. Gil-Rodríguez, and I. Manresa-Mallol, "How to Design a Mobile Learning Environment Recommendations Based on Student Perceptions," in 6th International Conference on Computer Supported Education CSEDU 2014, Barcelona, Spain, April 2014, vol. 3, pp.145-152.

[8] F. Kapp, I. Braun, H. Körndle, and A. Schill, "Metacognitive Support in University Lectures Provided via Mobile Devices How to Help Students to Regulate Their Learning Process during a 90-minute Class," in 6th International Conference on Computer Supported Education CSEDU 2014, Barcelona, Spain, April 2014, vol. 3, pp.194-199.

[9] S. Hadjerrouit, "Wiki-mediated collaborative writing in teacher education assessing three years of experiences and influencing factors," in Proceedings of the 6th International Conference on Computer Supported Education CSEDU 2014, Barcelona, Spain, April 2014, vol. 1, pp.5-14.

[10] I. Ajzen, "The theory of planned behavior," Organizational Behavior and Human Decision Processes., vol. 50, no. 2, pp. 179-211, 1991.

[11] V. Venkatesh, M. Morris, G. Davis, and F. Davis, "User Acceptance of Information Technology: Toward a Unified View.," Mis Quarterly, vol. 27, no. 3, pp. 425-478., 2003.

[12] D. Peng and F. Lai, "Using partial least squares in operations management research: A practical guideline and summary of past research.," Journal of Operations Management, vol. 30, pp. 467-480, 2012.

[13] E. Biddis and T. Chau, "Multivariate prediction of upper limb prosthesis acceptance or rejection," Disability and Rehabilitation: Assistive Technology, vol. 3, no. 4, pp. 181-192, 2008.

[14] P. Ifinedo, "Technology acceptance by health professionals in Canada: An analysis with a modified UTAUT model," in 45th Hawaii International Conference on System Sciences, Grand Wailea, January 2012, pp. 2937-2946.

[15] W. Chin, B. Marcolin, and P. Newsted, "A partial Least Squares Latent Variable modeling approach for measuring interactions effects: Results from a monte Carlo siulation study and an electronic-main emition / adotion study.," Information systems research, vol. 14, no. 2, pp. 190-217, 2003.

[16] W. Chin, "The partial least squares approach to structural equation modeling," in Modern Methods for Business Research, Mahwah, NJ, Lawrence Brlbaum Associates, 1998, p. 295–336..

[17] J. Hair, G. Hult, C. Ringle, and M. Sarstedt, A primer on partial least squares structural equation modelling (PLS-SEM), L.A.: Sage, 2014.

[18] A. Bere, "Exploring Determinants for Mobile Learning User Acceptance and Use: An Application of UTAUT," in 2014 11th International Conference on Information Technology: New Generations, Las Vegas, 2014.

[19] C. Davis and H. Goodman, "Virtual Communities of Practice in Social Group Work Education," Social Work with Groups, vol. 37, no. 1, pp. 85-95, 2014.

Web Service Selection Based on Integrated QoS Assesment

Olga Georgieva, Dessislava Petrova-Antonova
Department of Software Engineering
FMI, Sofia University “St. Kl. Ohridski”
Sofia, Bulgaria
e-mail: o.georgieva@fmi.uni-sofia.bg, d.petrova@fmi.uni-sofia.bg

***Abstract*—Currently, web services are the preferred technology for implementation of distributed applications that follow a Service-Oriented Architecture. The promise of reduced cost and time for enabling a large range of company-wide business processes shifts the research focus towards reasoning about web service selection. This paper proposes an approach facilitating clients to select a web service among several ones with the same functionality based on their quality of service (QoS) properties. It provides a mechanism for integrated QoS assessment of web services taking into account all measured QoS properties of the client’s interest. The method estimates the strength of the mutual dependency of the QoS properties using a data set analysis of QoS values accumulated during the web service invocation. The approach is summarized in a step-by-step assessment procedure and is proved through a real web service selection scenario.**

Keywords-Web services; Quality of service; Web service selection; Theory of fuzzy sets; Probability theory.

I. Introduction

Nowadays, the Service-Oriented Architecture is the most commonly used paradigm for development of distributed software systems, especially in the context of cloud computing. Since the web services are one of the fundamental technologies used in the cloud, the problems related to their quality remain of primary importance. Thus, many research efforts are still focused on the development of new approaches for quality of service (QoS) assessment.

A. Problem formulation

The web service selection is a challenging problem especially when different web services provide equal functionality [8]. The clients aim to use the web services having the best quality, but the selection procedure meets several difficulties as follows:

- The behavior of the web services with respect to QoS is difficult to be predicted. It depends on various factors such as availability of the network connection or corresponding application server, the number of simultaneous invocations and so on.
- The client's choice is a result of subjectivity of its own understanding about the desired quality of the offered resources [12].
- The web service quality is affected simultaneously by different QoS properties that often are inconsistent. For example, increasing the availability of a given web service leads to increasing of its cost, which is not reasonable for the clients with limited budget.

B. Current State of the Art

The recently introduced concept of web service quality estimation is based on the probability of the measured quality data and fuzzy sets (FSs) to account for the client preferences [1], [2], [3]. In [15] a web service selection approach based on the weight of client’s satisfaction from QoS properties is proposed. Its main drawback is the inability to ensure that the service recommending algorithm is open, fair and trustworthy. Also, only measurable QoS properties are considered. In contrast, a web service selection mechanism proposed in [16] deals with all types of QoS properties expressing those that are not measurable in terms of integer values. The QoS-based selection model described in [17], [18] determines the overall web service quality from a weighted sum of the normalized values of QoS properties. A drawback of the model is that it considers only measurable QoS properties that can be directly monitored. Also, the normalization of the values of QoS properties in the interval of [0,1] leads to losing valuable information. The QoS-aware selection method proposed in [19] performs credibility evaluation of QoS properties that are classified in two categories, i.e. negotiable and nonnegotiable. The values of nonnegotiable QoS properties are obtained from historical records of web service execution and cannot be modified by the provider. The negotiable QoS properties can be changed according to the client’s requirements. The web service selection algorithm presented in [20] is based on quantitative QoS prediction method applied to a dynamic environment. In [21] a collaborative filtering based approach is designed. It predicts QoS of unused web services taking into account the similarity in clients experiences. Such similarity is considered also in the service selecting model proposed in [22].

The disadvantages of the approaches presented above can be summarized as follows:

- **Applicability to only measurable QoS properties.** Most of the approaches do not take into account unmeasurable QoS properties.
- **Subjectivity of the performed QoS estimation.** Some approaches are based on the ratings that are specified by clients after web service consumption.

Others use values of QoS properties that are claimed by the web service providers.

- **Applicability to single QoS property.** A significant number of the proposed approaches can be applied in cases when clients are interested in one QoS property in particular. In practice, the web service selection procedure requires consideration of several QoS characteristics.

C. *Research Objectives*

In our previous work [6] we propose an algorithm that enables to compare the quality of several web services with equal functionality according to a single nonfunctional characteristic. It was implemented in a software tool for automated web service selection [14]. However, real practice shows that the service selection usually needs to account for more than one QoS property. An easily applied new method that calculates the level of satisfaction of several jointly estimated quality properties of a particular web service gives a theoretical frame able to assess the strength of the chosen nonfunctional characteristics of a web service [13].

The web service selection approach introduced in this paper is based on the assessed strength of the multiple QoS properties. It calculates the level of satisfaction of several jointly estimated QoS properties for each interested web service using the data collected during the web service invocation. The approach compares the obtained strength level in order to advise the client on the web service with the best quality. It is summarized in an easily implementable algorithm that is proven through a real web service selection scenario.

The rest of the paper is organized as follows. Section II presents the proposed approach for integrated QoS assessment. Section III describes a case study of web service selection as a proof of concept. Section IV concludes the paper and gives directions for future work.

II. Integrated QoS Estimated Method

This section presents a new approach for QoS-enabled selection of web services. First, a theoretical background of the approach is introduced. Next, an integrated QoS assessment for web services is proposed. Finally, a step-by-step procedure for QoS-enabled selection of web services is described.

A. *Theoretical background*

The QoS properties of web services could be divided into groups, namely measurable and unmeasurable [10]. The QoS properties such as response time, availability and throughput are measurable, since they have numerical values. The unmeasurable QoS properties such as standard compliance, authentication method and data encryption cannot be directly measured in terms of numeric values. In contrast, the values of a given measurable QoS property could be easily obtained through monitoring of the web service behavior during operation invocation. They can be treated as values of a random discrete variable known as Probability Mass Function (PMF) [4]. If x is any possible value of a discrete random variable X, the PMF value of x, denoted by $p(x)$, is the probability of the event $\{X=x\}$ consisting of all outcomes that give rise to the value of X equal to x:

$$p(x) = P(\{X=x\}). \qquad (1)$$

The PMF is appropriate for QoS assessment of the web services regarding a single QoS property [11]. Unfortunately, it cannot be applied to the unmeasurable QoS properties. For instance, the QoS metrics that reflect the subjectively experienced quality as usability, efficiency, etc., are in fact the acceptable cumulative effect on client satisfaction of all imperfections affecting the web service [5]. These qualities could be only empirically assessed according to the subjectivity of the client's perception sensed during the web service usage. Usually, such assessment is not exactly defined as it implies some uncertainty. This type of uncertainty can be formalized using a powerful mathematical tool of the theory of fuzzy sets [7].

The concept of fuzzy set allows partial set membership rather than a crisp set membership. The level of belonging is formulated by a membership function, which numerically represents the degree to which a given element belongs to a fuzzy set. Formally, a fuzzy set is defined as follows: a fuzzy set A defined in a universe U is a set of ordered pairs, $A = \{(x, \mu_A(x)), x \in U, \mu_A(x) \in [0,1]\}$, where $\mu_A(x)$ is the degree of membership of the element x in A. Thus, the membership function is seen as a mapping $\mu_A(.)$: $x \rightarrow [0,1]$. As $\mu_A(x)$ approaches 1, the element x increasingly belongs to the fuzzy set A. Commonly, the shape of the membership function is identified as a standard function of triangular, trapezoidal, Gaussian or other type. Usually, it is determined by expert knowledge. However, there are different practical applications, which successfully apply techniques of exploratory data analysis for membership function identification. In such a case, $\mu(.)$ covers the intrinsic process uncertainty that leads to impossibility to find an accurate process description [7]. Since the information for a given QoS property is collected in a data set of measured values then a discrete membership function is applicable to the web service assessment problem. It is interpretable in the sense of probability distribution obtained by (1) [6].

Usually, several quality properties are significant for the web service selection problem. Often they have contradictory effect on the whole service behavior. By improving the quality of one nonfunctional characteristic the quality in the sense of another could decrease. For instance, when web service reliability and safety are increased, its response time could be worsted. Therefore, the joint assessment of several quality properties is essential for the QoS-enabled web service selection.

B. *Integrated QoS estimation*

Let us consider that we have m different web services S_j $j = 1,\ldots,m$ presenting the same functionality. Let us also assume that we have n number of quality properties x_i that determine the QoS properties according to which the web services have to be compared. If each quality property value is presented as a fuzzy set Q^j_i, $j = 1,\ldots, m$, $i = 1,\ldots,n$ as it discussed above, its membership function $\mu_{Q^j_i}(x_i)$ could be represented as PMF according to (1). In such a case, the joint relation of all interested web service QoS properties could be estimated via Cartesian product calculation [13]. The Cartesian product of the fuzzy sets Q^j_i defines a new fuzzy set Q^j of the cross product for the *j*-th service:

$$Q^j = Q^j_1 \times Q^j_2 \times \ldots \times Q^j_n. \qquad (2)$$

The FS Q^j is a set of all pairs that consists of a tuple of QoS properties' values $x^j = (x^j_1, \ldots, x^j_n)$ and its membership degree $\mu_Q^j(x^j)$. That membership degree represents the strength of the relationship of the QoS values of the respective tuple x^j. It is calculated as a minimum of the membership degrees of the constituent membership degree values:

$$\mu_{Q^j}(x^j) = \min_{i=1,n}(\mu_{Q^j_i}(x^j_i)), \text{ for each } j = 1,\ldots,m. \qquad (3)$$

The degree values obtained in (3) could serve as a QoS assessment and a web service comparison analysis. From a practical point of view, the larger membership degree shows more strength of the QoS. By maximizing the strength, we are able to select a web service, which provides maximal quality. For the QoS properties, whose values have to be minimized in the assessment procedure, the negation of the membership degree:

$$\neg\mu_{Q^j_i}(x^j_i) = 1 - \mu_{Q^j_i}(x^j_i) \qquad (4)$$

should be taken in the calculation of (2) and respectively of (3). Thus, at the final stage the largest $\mu_Q^j(x^j)$ value corresponds to the web service that has the best quality and shows the most preferable service behavior:

$$\mu_{Q\max} = \arg\max_{j=1,n}(\mu_{Q^j}(x^j)), \text{ for } j = 1,\ldots,m. \qquad (5)$$

C. *Integrated QoS selection method*

The theoretical statements presented above can be summarized in a procedure that provides an integrated QoS web services assessment. Let us assume that for a given web service, a number of *n* QoS properties x_i, $i=1,\ldots,n$ are of a particular interest. Let the candidate web services, namely those that satisfy functional requirements of the client, form the set *S*, where $S=\{S_1, S_2, \ldots, S_j \ldots S_m\}, j=1,\ldots,m.$

The steps of the proposed approach for integrated QoS assessment are as follows:

Step 1: Accumulate data for all interested QoS properties within a chosen time window.

Step 2: Represent each property of each service as a fuzzy set Q^j_i, $j=1,\ldots,m$, $i=1,\ldots,n$ corresponding to a membership function $\mu_Q{}^j{}_i(x^j{}_i)$.

Step 2 requires calculation of the PMF for each QoS property. In case of available data about the incidence of the QoS values it is processed according to (1) in order to obtain the probability and further, interpret the probabilities values as fuzzy membership degrees. For the properties whose values have to be minimized in the assessment procedure, the negation of the membership degree is performed according to (4).

Step 3: Estimate the Cartesian product of the fuzzy sets Q^j_i, $j=1,\ldots,m$, $i=1,\ldots,n$ for each web service according to (2) and (3).

Step 4 Apply (5) to select the highest strength that finds the best web service regarding the quality for the considered time window.

III. Method Validation

The feasibility of the proposed QoS-enabled approach for web service selection is proved by a sample case study. It covers four web services providing the same functionality of email validation. The web services are compared according to three QoS properties. They are throughput (TP), response time (RT) and number of completed requests (CR). The dataset with QoS values is obtained through load testing of the web services using the LoadUI tool [9]. The TP gives the data transfer rate and is typically measured in bytes per second. The RT gives the time in milliseconds taken to send a request and to receive a response from the web service. The CR provides information about the number of successful processed requests by the web service. Each web service receives 10 requests per second during test case execution. The total number of requests is 1220 and the test case execution time is approximately 20 minutes. This time interval is enough for QoS estimation of the chosen QoS properties, since the behavior of the web services for a longer period of time become repeatable and does not affect the proposed method for QoS estimation. The minimal (*RTmin, TPmin, CRmin*) and maximal values (*RTmax, TPmax, CRmax*) of the QoS properties as well as their maximum of the respective probability *max(P)* for all four web services are presented in Table I, Table II and Table III.

TABLE I. Response Time

Web service	**Response Time**		
	max(P)	***RTmin***	***RTmax***
webserviceex	0.0131	197,7	858.8
cdyne	0.0107	145,82	7308
postcodeanywhere	0.0164	62	383
serviceobjects.net	0.009	324,18	580

TABLE II. Throughput

Web service	**Throughput**		
	max(P)	***TPmin***	***TPmax***
webserviceex	0,6959	17379	77775
cdyne	0,7270	0	922240
postcodeanywhere	0,8885	3183	60477
serviceobjects.net	0,5205	209493	302601

TABLE III. COMPLETED REQUESTS

Web service	Completed requests		
	max(P)	***CRmin***	***CRmax***
webserviceex	0,7057	3	20
cdyne	0,7279	0	44
postcodeanywhere	0,8885	1	19
serviceobjects.net	0,5279	9	13

The PMFs calculated for the RT are graphically presented in Figure 1. The maximum of PMFs for all web services is obtained for RT lower than 400 milliseconds. The “cdyne” web service has a big diversity of singly measured values of RT. Its RT varies in a large time interval (see the minimum and maximum value in Table I), but most often is not higher than 200 milliseconds. The X-axis values of the Figure 1 are limited to 2000 and by that increasing the readability and showing better the RT values that are most frequently obtained during web services testing.

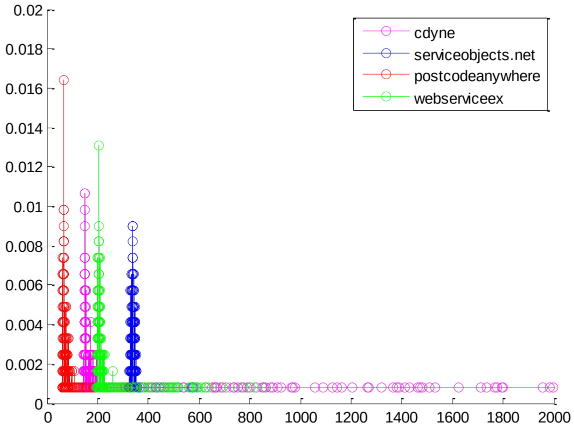


Figure 1. PMFs of Response time.

The PMFs calculated for the TP are graphically presented in Figure 2.

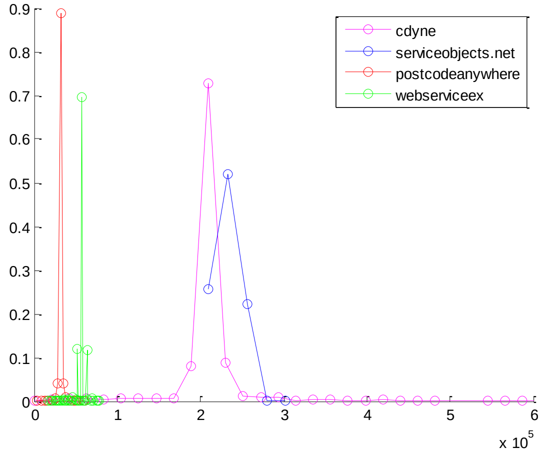


Figure 2. PMFs of Throughput.

The “cdyne” web service has maximum value for the TP 922240 bps (Table II). In order to gain a better readiness the X-axis values of the Figure 2 are limited to 600000. The similar refinement is made regarding the Figure 3 where the X-axis values are limited to 25.

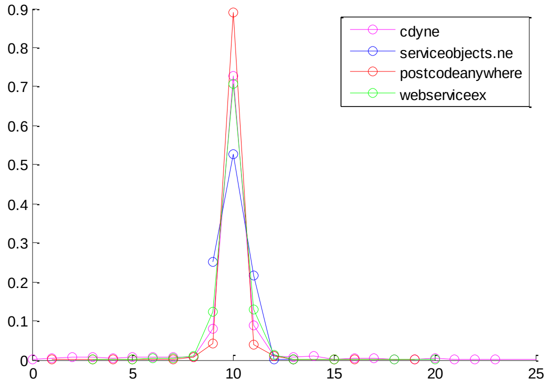


Figure 3. PMFs of Completed requests.

The obtained dataset is used for integrated assessment of the web services’ quality. Note that the web services clients expect as higher values as possible for TP and CR. In contrast, regarding the RT the quality is higher when its values are smaller. That is why the PMF of RT is not directly used in calculation procedure, since it needs to be negated according to (4) to obtain the membership degrees of the corresponding fuzzy set Q_{RT}. The PMF of TP and CR are treated as fuzzy sets Q_{TP} and Q_{CR} having the membership degrees that are used directly in the assessment algorithm.

TABLE IV. QOS ASSESMENT

Web service	μ_{Qmax}
webserviceex	0.6959
cdyne	0.7270
postcodeanywhere	0.8885
serviseobjects.net	0.5205

The Cartesian product $Q = Q_{TP} \times Q_{RT} \times Q_{CR}$ of the fuzzy sets related to the QoS properties of interest is calculated according to (2) and (3). The results from the QoS estimation are shown in Table IV. As it was underlined, the membership grade defines the strength of the relationship between the values of QoS properties for each triple. The highest value of 0,8885 is obtained for “postcodeanywhere” web service. It provides the highest quality and is recommended to the client as the best web service for email validation.

IV. CONCLUSION

The proposed approach provides a procedure that assesses the strength of several QoS properties of different web services with equal functionality. It is based on a solid theoretical frame for joint assessment of the QoS properties of interest. The approach calculates the level of satisfaction of the QoS properties using the fuzzy sets description and their cross product in order to find the strength of the relationship between the QoS properties. This is proved through a real web service selection scenario.

Since the proposed approach is based on data obtained through monitoring, it could be realized in a fully automated manner. Further, it uses a theoretical basis that enables to be considered not only measured but also unmeasured QoS properties. The future work includes comparison of the proposed method with other approaches in case of equal web service scenarios and to evaluate the method's effectiveness with larger QoS datasets.

ACKNOWLEDGMENT

The work was supported by the project No 24/2015 sponsored by the Science Fund of Sofia University "Kl. Ohridski", Bulgaria.

REFERENCES

[1] K. F. Abbaci et al., "Selecting and Ranking Business Processes with Preferences: An Approach Based on Fuzzy Sets," Lecture Notes in Computer Science, vol. 7044, Springer-Verlag Berlin Heidelberg, pp. 38-55, 2011.

[2] S. Amdouni et al., "Answering Fuzzy Preference Queries over Data Web Services," Lecture Notes in Computer Science, vol. 7387, Springer-Verlag Berlin Heidelberg, pp. 456-460, 2012.

[3] K. Benouaret et al., "Selecting Skyline Web Services from Uncertain QoS," Proc. Int. Conf. on Services Computing, Honolulu, Hawaii, USA, 2012, pp. 523-530.

[4] J. Devore, Probability and Statistics for Engineering and the Sciences. Brooks/Cole Publisher, Belmont, USA, 2008.

[5] L. Franken, Quality of Service Management: A Model-Based Approach. PhD thesis, Centre for Telematics and Information Technology, 1996.

[6] O. Georgieva and D. Petrova-Antonova, "QoS-aware Web Service Selection Accounting for Uncertain Constraints," Proc. 40th IEEE Euromicro Conference on SEAA, Verona, Italy, 2014, pp. 174-177.

[7] G. Klir and B. Yuan. Fuzzy sets and fuzzy logic: theory and applications, Prentice Hall, USA, 1995.

[8] M. Li et al., "An Adaptive Web Services Selection Method Based on the QoS Prediction Mechanism," Proc. Int. Joint Conf. on Web Intelligence and Intelligent Agent Technology – Workshops, 2009, pp. 395-402.

[9] LoadUI tool. http://www.loadui.org/ (May, 2015).

[10] V. Tran et al., "A new QoS ontology and its QoS-based ranking algorithm for Web services," In Simulation Modelling Practice and Theory, vol. 17, no. 8, pp. 1378-1398, 2009.

[11] D. Petrova-Antonova, "A QoS Aware Approach for Web Service Selection Based on Probability Evaluation," Proc. the IADIS International Conference on Informatics 2011, Rome, Italy, July 2011, pp. 43-50.

[12] P. Xiong and Y. Fan, "QoS-aware Web Service Selection by a Synthetic Weight," Proc. Int. Conf. on Fuzzy Systems and Knowledge Discovery. Haikou, Hainan, China, 2007, pp. 632-637.

[13] O. Georgieva, "Joint assessment of Software Service Quality Properties," Proc. 13th International Conference e-Society, Funchal, Portugal, March 2015, pp.316-315.

[14] D. Petrova-Antonova, N. Hristova and O. Georgieva, "RecSS: automation of QoS-aware web service selection," Proc. the 15th International Conference on Computer Systems and Technologies, Rousse, Bulgaria, June 2014, pp. 256-263, doi 10.1145/2659532.2659632.

[15] S. Li et al., "A Mechanism for Web Service Selection and Recommendation Based on Multi-QoS Constraints," Proc. 6th World Congress on Services. Miami, USA, 2010, pp. 221-228.

[16] D. D'Mello et al., "A QoS Broker Based Architecture for Dynamic Web Service Selection," Proc. Asia Int. Conf. on Modelling & Simulation. Kuala Lumpur, Malaysia, 2008, pp. 101-106.

[17] L. Sha, et al, "A QoS based Web Service Selection Model," Proc. Int. Forum on Inf. Technology and Applications. Chengdu, China, 2009, pp. 353-356.

[18] R.J.R. Raj and T. Sasipraba, "Web service selection based on QoS Constraints," Proc. 2nd IEEE Int. Conf. on Trendz in Information Sciences & Computing, Chennai, 2010, pp. 156-162.

[19] L. Qi, et al., "A QoS-Aware Web Service Selection Method Based on Credibility Evaluation," Proc. Int. Conf. on High Performance Computing and Communications. Melbourne, Australia, 2010, pp. 471-476.

[20] M. Li et al., "An Adaptive Web Services Selection Method Based on the QoS Prediction Mechanism," Proc. Int. Joint Conf. on Web Intelligence and Intelligent Agent Technology – Workshops, 2009, pp. 395-402.

[21] L. Shao et al., "Personalized QoS Prediction for Web Services via Collaborative Filtering," Proc. Int. conf. on Web Services. Salt Lake City, USA, pp. 439-446, 2007.

[22] S. Wang and H. Chen, "A Web Service Selecting Model Based on Measurable QoS Attributes of Client-Side," Proc. Int. Conf. on Computer Science and Software Engineering. Wuhan, China, 2008, pp. 385-389.

Does the Integration of the Concept of Rapid Instructional Design in Project Management Approaches Support the Efficient Realisation of E-learning Projects?

Michał Kuciapski
Department of Business Informatics
University of Gdansk
Gdansk, Poland
e-mail: m.kuciapski@univ.gda.pl

Abstract— **The development and implementation of interactive e-learning courses is highly time-consuming and expensive. A question, therefore, arises whether it is possible to develop an e-learning program at a high level without incurring significant financial resources and the need to create a project team consisting of a wide range of specialists. In this respect, Rapid Instructional Design (RID) might have a useful role to play in that it focuses on a reduction of the cost and time involved in developing e-learning courses, while ensuring a reasonable level of quality. The purpose of this article is to present an adaptation of the concept of RID for proposing and verifying a practical approach to the implementation of e-learning. The article starts with an analysis of the problem of how to create an e-learning course with limited time, financial and human resources. Then, the article continues with a critical analysis of the RID concept. The third section of the article outlines an approach to the management of e-learning projects based on RID, taking into consideration the problems faced when developing e-learning courses. The fourth and fifth sections present the methodology and verification results of the developed approach which answer the question considered during the research - does the integration of the concept of Rapid Instructional Design support the efficient realization of e-learning projects? The last section of the article contains a critical discussion of the possible applications of the adapted concept of RID in various types of projects, based on the research of its implementation in two e-learning projects.**

Keywords- rapid instructional design; e-learning; higher education; project management; course development; quality management.

I. INTRODUCTION

Despite the significant development of distance learning in the past few years, e-learning (network learning) and m-learning (mobile learning), which are the newest forms of d-learning (distance learning), are still under-utilized [22]. This, in turn, limits the adaptation of distance learning on a large scale. The use of even more innovative Web 2.0 e-learning tools does not change the conclusion that it is the teaching material that still plays the key role. The challenges of modern academic teaching and the training market in conjunction with the rapid development of Information and Communication Technologies (ICT) have led to more demanding quality requirements regarding e-learning [15]. An important indicator of the attractiveness of an e-learning course, in addition to the content, it is also the quality of the presentation of the educational material, often associated with the level of interactivity and multimediality [16]. In this regard, it is important to develop and integrate components, such as: illustrations, interactive graphic, movies, animations and simulations [24]. These cannot be simply "art for art's sake", but must serve the key objective regarding the multimedia content and interactivity of e-learning courses, that is to improve the efficiency of learning [21]. This can be achieved when the multimedia objects involved in e-learning courses support a faster acquisition of knowledge and skill development [20]. The interactivity of e-learning material can also help to maintain a high level of commitment from the participants of e-learning courses.

The professional production of e-learning materials consistent with the assumptions outlined above is extremely time-consuming and expensive [25]. In order for the financial resources incurred in converting static teaching material into multimedia and interactive versions to be spent effectively, it is necessary to create a project team including a range of professionals [1], as indicated in Table 1. In addition, as shown in Figure 1a, the preparation and implementation of high quality e-learning courses requires the management of many parallel processes. This requires a commitment from project managers with relevant experience in e-learning initiatives [6]. Therefore, universities and training companies often have difficulty in preparing e-learning programs of high quality or on an appropriately large scale.

Therefore, it is particularly relevant to seek approaches which may enable universities and companies to implement e-learning, but within the strict financial and time constraints faced by many organizations. These objectives are considered to be key in terms of the RID. Referencing RID to other alternative models for the development of distance learning, such as the Classroom Oriented Model [3], Product Oriented Model [2] or the System Oriented Model [10], it focuses more widely on practical aspects. The RID concept has been touched upon by only a few authors, so remains under-studied.

Outlined in the current point of the article premises are the basis for the formulation of two hypotheses:

1. The adaptation of RID in approaches to the project management of e-learning course development supports the creation of e-learning courses at an acceptable level

of quality while significantly reducing the time and cost of project realization relative to traditional approaches.

2. The use of RID is a useful alternative in the development of e-learning courses, compared to preparing a simple e-learning programme based on static documents or expensive multimedia and interactive e-learning materials.

The verification results of the stated hypotheses will answer the question posed in the title of the article – does the integration of the concept of Rapid Instructional Design in project management approaches support the efficient realization of e-learning projects?

The second section of the article contains a review of relevant literature and critical analysis of the concept of RID. The third section of the paper presents a proposed solution - an approach to the project management of e-learning course development that integrates the concept of RID to significantly reduce the time and cost of implementing e-learning. The validation methodology of the proposed approach is outlined in the fourth section of the paper. The fifth section of the paper presents the validation results of the proposed solution that answer the research question - Does the integration of the concept of Rapid Instructional Design in project management approaches support the efficient realization of e-learning projects? The paper finishes with a discussion and conclusion.

II. Background and related research

According to Clark [7], a key step in the preparation of an e-learning course is its design. An approach that can be used in a number of e-learning initiatives, in order to speed up the process, is the concept of Rapid Instructional Design. The direct meaning of the word ‘rapid’ does not fully reflect the specifics of RID. This concept, analogous to extreme software engineering methodologies, aims at shortening the development time and reducing the role of documentation. The RID concept was introduced by Thiagarajan [23], who devised the key objectives of RID to replace the traditional model of e-learning design - instructional design system (ISD) - with: the consistent creation of training packages, an acceleration of the design process, and use of appropriate shortcuts, borrowings and omissions from the ISD model. The consistent creation of training packages is based on ‘just-in-time’ method, for the daily delivery of learning packages, unlike the conventional ISD model that assumes the sequential realization of these processes. In this respect [23] identified 10 strategies with 20 directives, among which the most important are: the use of existing learning resources, utilization of templates, integration of tools for the acceleration of ISD processes as well as a better and wider use of human resources.

The RID concept was extended in [20]. Accordingly, projects based on RID, reductions in the duration and cost of developing an e-learning programme are achieved by simplifying, wherever possible, the standard activities that make up the design process like: analysis of the material, choosing methods of adaptation, preparation of instructional scripts, as well as the production and integration of the developed components of e-learning courses [19]. In particular, the instructional design should be limited in scope. Instructional design by [9] is a systematic approach to the design of teaching instructions as attractive learning scenarios, according to the customers’ needs and with the possibility for adaptation as a multimedia e-learning course.

Despite its name, RID implies the accelerated execution of all processes, not just design. This is achieved by carrying out specific steps, while often omitting certain tasks. Support from RID in this area takes the form of a series of best practices, with the most significant being [23]:

- a needs analysis based on the materials immediately available or soon to be, such as existing documents, instead of conducting interviews, for example;
- the production of ready-made templates to facilitate the preparation of multimedia objects and activities;
- the use of a reviewing system of the e-learning course specifications by a team of professionals with assigned roles to enable the faster execution of the evaluation;
- assessment and review of the e-learning course based on representative testing, or a pilot, allowing the verification of the e-learning course quality by some of the participants before running it.

The above list [13] may be expanded by adding the training of the authors in the application of technologies that enable the independent and rapid production of selected types of multimedia objects. Certain RID practices do not consider the important role that e-learning and ICT can play in facilitating the preparation of e-learning courses. Their potential caught the attention of [18] and [14] who distinguished a number of principles among which the most important ones are:

- the use of dedicated software to support the acceleration of the process of design and production of e-learning course components;
- the use of templates for: multimedia objects, activities, and other elements, wherever possible;
- the use of Reusable Learning Objects (RLO) and components;
- the use of ready-made commercially available e-learning courses or teaching material;
- the use of existing solutions purchased from external suppliers;
- the use of tools to facilitate the needs analysis process;
- the use of Training Management Systems (TMS) to facilitate the preparation of e-learning courses with integrated tools.

It should be noted that RID takes the form of a series of best practice recommendations rather than a precise methodology and is not based on a specific approach to project management. This allows the principles of RID to be adapted according to different models and methodologies of e-learning project management.

III. Solution proposal

A concept regarding the practical implementation of the principles of RID was developed for projects such as CTF

[4] and Case Simulator [5], carried out in 2012-2014. According to the principles set out primarily by Thiagarajan [23] and Piskurich [18] [19], their approach for the project management of e-learning (Figure 1b) includes:

1. A modification of the processes existing in the classical model of an e-learning project. The basic model is considered to be the ADDIE (Analyze, Design, Develop, Implement, Evaluate), which is very general in character. The author used developed model (Figure 1a), which is compatible with the ADDIE model, and expanded it with popular approaches in relevant books and articles [11] [8] [17].
2. The wide use of software to support the acceleration of design and production processes.
3. Process support via the use of already available templates of components.
4. The omission of a range of activities during processes carried out at the conceptual and executive stages.

The general model for adopting the concept of RID for project preparation and implementation of e-learning offer compared to a traditional approach is shown in Figure 1.

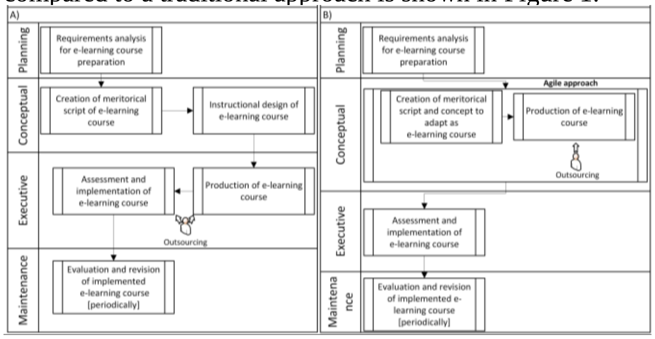

A)
Planning
Requirements analysis for e-learning course preparation
Conceptual
Creation of meritorical script of e-learning course
Instructional design of e-learning course
Executive
Assessment and implementation of e-learning course
Production of e-learning course
Outsourcing
Maintenance
Evaluation and revision of implemented e-learning course [periodically]
B)
Planning
Requirements analysis for e-learning course preparation
Agile approach
Conceptual
Creation of meritorical script and concept to adapt as e-learning course
Production of e-learning course
Outsourcing
Executive
Assessment and implementation of e-learning course
Maintena nce
Evaluation and revision of implemented e-learning course [periodically]

Figure 1. Traditional approach of e-learning project management (a), The approach to e-learning project management based on the RID concept (b).

As shown in Figure 1, the modification of the traditional model of e-learning course development, in order to bring in the concept of RID, primarily concerned the design and production processes, which are combined in a single process. Thus, the main processes to be shortened and simplified are the design and production processes of e-learning course development (Table 1). These are by far the most time consuming and cost intensive processes [12]. Other processes are carried out without any significant changes (Figure 1). In the developed model, the following elements of RID were adapted:

- a requirement analysis based on the materials immediately or quickly available instead of, for example, conducting interviews;
- a system of reusable components, through the preparation and implementation of e-learning in accordance with Sharable Content Object Reference Model (SCORM) standard and the production of similar multimedia objects based on templates;
- the direct design of the e-learning course structure with the use of authoring tools;
- the production of multimedia materials using tools supporting rapid animations and interactions generation based on preliminary concepts, to reduce the amount of multimedia objects' specification.

A list of the differences between the traditional approach and adaptation of the RID concept is shown in Table 1.

TABLE I. COMPARISON OF TRADITIONAL AND RID BASED APPROACHES FOR E-LEARNING COURSES DEVELOPMENT

Process	Implementation (according to approach)		Executor	
	Traditional	***RID***	***Traditional***	***RID***
Requirements Analysis	Detailed analysis of the environment and training needs along with the target group.	Basic analysis of the environment and training needs along with the target group.	Instructional designer, author, expert	Author
Script creation	Preparation of a script along with evaluation elements as well as an initial material adaptation concept.	Preparation of a script along with evaluation elements as well as a material adaptation concept.	Author, instructional designer, reviewer	Author, designer
Design	Preparation of comprehensive specifications of multimedia objects as well as training structure.	One process for creating e-learning course. Direct design in authoring tool and other auxiliaries of the e-learning structure as well as multimedia objects based on their concept. Simplification of multimedia objects. Preparation of implementation packages.	Instructional designer	Designer
Production	Production of multimedia items. Validation and integration of components. Preparation of implementation packages.		Instructional designer, Multimedia team (graphic designer, audio-video specialist and others).	
Assessment and implementation	Final evaluation of the e-learning course and implementation of updates.	Same as in the traditional approach.	Author, instructional designer, tutor, multimedia team, platform administrator	Author, designer, tutor.
Evaluation and revision	Evaluation of the e-learning course content and attractiveness and proper changes.	Same as in the traditional approach.	Author, instructional designer, tutor, multimedia team.	Author, designer, tutor.

According to Table 1, the omission and simplification of a range of activities is related to the processes of requirement analysis, design and production of the e-learning course. This reduces the time and costs of the last two processes mainly. At the same time, the approach assumes an expansion of the design process where the author creates the main script and the concept to adapt it as an e-learning course with multimedia objects. Significantly, Table 1 reveals a significant reduction in the number of members required for a project team, as selected activities are delegated to:

- the author – conducting the requirement analysis and preparing the concept for the adaptation of material for e-learning course. These tasks are traditionally done by instructional designer.
- the instructional designer - the production of multimedia objects, integrating the components that comprise an e-learning course. These activities in classical approaches are carried out by the multimedia team.

The project team was limited in members, in particular by the omission of contractors responsible for the production of multimedia. This is possible thanks to the decision not to make any technical specifications for the multimedia materials and to reduce their complexity. In this way, the media production tasks are assigned solely to the designer.

Unlike the concept of RID, the developed approach to the project management of e-learning course development did not include the training of authors in terms of utilizing tools for the self-production of selected types of multimedia objects. This was due to the specific nature of the CTF and the CaseSimulator projects. E-learning courses were developed as small solutions, based on 20-30 pages of script written by several authors for each e-learning course. A considerable amount of time would be necessary to conduct training for the authors. Moreover, the time spent on training the authors would be longer than the time required to produce the e-learning courses. Accordingly, operations were carried out exclusively by instructional designers.

Entrusting the production of e-learning courses entirely to the designers would not be possible without the use of appropriate technology for the production of courses and pre-built templates previously prepared by professionals, such as web-masters, graphic designers and SCORM programmers. In this regard, a sophisticated authoring tool, supporting basic and intermediate animations and interactions creation, was chosen to offer ample opportunities to produce multimedia materials. Mechanisms for specifying transformations provided by the application enabled the immediate development of basic animations for multimedia objects. This eliminated the need to use graphic designers and programmers for the multimedia production and simplified the development process of e-learning materials, helping to reduce the associated costs. The financial resources required were also reduced by excluding the preparation of special graphics made by graphic designers by using alternatives available online for free or by buying a repository license.

IV. Research Methodology

The approach developed based on the concept of RID was used during the CTF project, in the years 2012-2013. Twelve specialized e-learning courses were prepared and offered in the area of economics, particularly in entrepreneurship. Courses were available to employees of small and medium-sized enterprises on the catching.ug.edu.pl platform. Over 600 employees participated in the courses. This project management approach was also carried out during the CaseSimulator project. In 2013-2014, 3 e-learning programs were prepared and offered to students, presenting the use of an information system for simulating the running of a business.

The factors that were used to validate the project management approach based on the concept of RID were as follows: the time and cost incurred for developing e-learning courses and the quality of the e-learning courses. The analysis of the time and costs of the developed approach was carried out in comparison with traditional solutions. Traditional verification approaches were made in 2008-2011 during the project "The Implementation of Modern Education in the University of Gdansk", when 6 fully interactive multimedia and e-learning courses for students and academic staff were devised. Time and cost monitoring for the RID-integrated project management approach was performed during the CTF and Case Simulator projects, realized in 2012-2014.

The quality of e-learning courses prepared according to traditional and RID-integrated project management approaches was assessed with the use questionnaires and interviews with authors (20) and participants (943). The results were compared with the assessment of e-learning courses developed with the use of a traditional project management approach.

V. Research Results

E-learning courses prepared according to a traditional approach were developed on the basis of 170 pages of material on average, as opposed to an average of 20 pages in the case of the RID- integrated approach. Therefore, Table 2 shows data for the elaboration and implementation of 8.5 RID-based e-learning courses. The cost calculations do not include hardware, software or office supplies.

TABLE II. Comparison of the time and cost of e-learning course development depending on the approach

Process	Time (working days)		Cost (USD)	
	Traditional	***RID***	***Traditional***	***RID***
Requirements analysis	Omitted			
Script preparation	Omitted			
Design	88	10.5	7330	1700
Production	66		7500	
Assessment and implementation	15	3	1139	558
Evaluation and revision	Omitted			
Total	169	13.5	15969	2258

Table 2 clearly indicates a much shorter time and lower costs for e-learning courses developed with a RID-based project management approach compared with traditional methods. The time and cost of the requirements analysis process and the creation of the script are not dependent on the approach, were not relevant to the comparison and were therefore omitted. The difference in the duration and cost of the assessment and implementation of e-learning exists due to the time spent on course verification, improvements or updates. E-learning courses prepared in accordance with the developed approach were characterized by a simplification of the concept, and therefore of the multimedia objects. The preparation of an e-learning course by an instructional designer rather than a multimedia team. Therefore, as a result of these simplifications, the multimedia materials experienced significantly fewer bugs, with less time and money for fixing. The duration and cost of the course implementation was the same regardless of used approach.

For a precise comparison of the time and costs required to prepare and implement e-learning courses according to the approach, these values were calculated for the script page as a main comparison unit, and are presented in Table 3. When calculating the cost of e-learning projects depending on the approach, costs that are similar regardless of the use or non-use of RID were disregarded: requirement analysis and the creation of the script, as well as hardware and software.

TABLE III. COMPARING THE TIME AND COSTS FOR E-LEARNING COURSE DEVELOPMENT BETWEEN A TRADITIONAL AND AN RID-BASED PROJECT MANAGEMENT APPROACH

Factor	Traditional	RID	Relation (%)
Time (days/page)	1	0.075	7,5
Costs (USD/page)	93.95	13.28	14

According to Table 3, the adaptation and implementation of a page of script as the webpage of an e-learning course, using an approach based on the concept of RID was 13.33 times faster and 7.14 times cheaper than the traditional method. This indicates that the use of a methodology based on RID can enable the development of e-learning courses with significantly limited time and financial resources. Thus, the utilization of the developed approach gives important opportunities to universities and companies in terms of the adaptation of traditional materials for e-learning courses, and thus the preparation of a wide e-learning programme.

The key factor in the assessment of the developed model is the verification of the quality of e-learning courses developed with the use of this approach compared with e-learning courses produced and implemented by traditional methods. Analysis showed that e-learning courses developed with the use of a traditional approach and those which are RID-integrated have a similar visual quality. A more important issue is the reception of e-learning courses by the script authors and participants. In this regard, a verification of traditional approaches was made in 2008-2011 during the project "The Implementation of Modern Education at the University of Gdansk". The validation of the approach developed to integrate the concept of RID was carried out for the previously mentioned CaseSimulator and CTF projects. A summary of the quality assessment - resulting from questionnaires and interviews - for the respective project management approach is shown in Table 4.

TABLE IV. ASSESSMENT OF THE QUALITY OF E-LEARNING COURSES PREPARED ACCORDING TO TRADITIONAL AND RID-INTEGRATED APPROACHES

	Traditional (6 e-learning courses)		***RID (15 e-learning courses)***	
	Amount	*Evaluation score*	*Amount*	*Evaluation score*
Authors	8	High evaluation score for the quality of the e-learning courses. A quantitive analysis revealed that 12,75% of multimedia items contained significant errors compared with 78,5% of items with few or no errors.	12	No reservations regarding the quality of the e-learning courses. Acceptation of the quality of educational and visual content as well as the functionality of the e-learning courses – only quality research with interviews.
Participants	183	Very high evaluation score for the quality of the e-learning courses. A quantitive survey was carried out (102 answers received) in which the attractiveness of the presentation of materials was scored highly or very highly (86,5%) compared with the 0% of students who rated the materials as unattractive.	760	No reservations regarding the quality of the e-learning programme. Acceptance of quality of the educational and visual content as well as the functionality of the e-learning courses - quality research with interviews. No quantitive surveys.

Table 4 indicates the quality of e-learning courses prepared using a traditional approach compared with one integrating the concept of RID. E-learning courses prepared using a conventional approach were given a higher rating.

Taking into account the validation results of the approach developed according to the concept of RID, in terms of factors such as the time and cost of preparation as well as the quality assessment of the e-learning courses, the two stated hypotheses were confirmed:

1. The adaptation of RID in approaches to the project management of e-learning course development supports the creation of e-learning courses at an acceptable level of quality while significantly reducing the time and cost of project realization relative to traditional approaches.
2. The use of RID is a useful alternative in the development of e-learning courses, compared to preparing a simple e-learning programme based on static documents or expensive multimedia and interactive e-learning materials.

The confirmation of the two stated hypotheses positively answers the question posed in the title of the article - does the integration of the concept of Rapid Instructional Design in project management approaches support the efficient realization of e-learning projects?

VI. DISCUSSION

The use of the developed approach to e-learning project management, as well as the validation results of this approach, enable a critical evaluation of the RID concept. First and foremost, it is apparent that the RID-based solutions are usable when the multimedia materials to be produced will involve a low or moderate complexity of animations and interactions. In addition, in approaches to e-learning project management based on RID, it is desirable for the instructional designer to have basic skills in the field of graphics and animation programming.

The reason is that RID primarily concentrates on reducing expenditure on the design and production of multimedia objects, and thus the multimedia team. When the necessity arises to prepare complex animations with complicated interactions involving the use of audio and video, such as simulations, the concept of RID is highly difficult to apply. In such a situation it is of the utmost importance to develop a detailed instructional design specification for multimedia objects and create a multimedia team. However, even in such a scenario, the concept of RID can be partially applied, where the traditional production process will refer only to selected complex multimedia materials. This will continue to reduce the time and costs of projects for e-learning course development.

One of the principles of RID is to delegate tasks. In this regard, practical experience shows the possibility of offloading design tasks to the authors, and production to instructional designers. However, delegation per se does not reduce the cost and duration of the project. Moreover, in more complex projects, it may turn out that the authors and designers do not have sufficient skills and are less effective in their work than a multimedia team. It should be noted that in such a situation, the opposite effect is achieved in terms of reducing expenditure.

The key principle of RID assumes the simplification of the design and production processes of e-learning courses preparation, with an impact on the multimediality and interactivity of learning material. It is natural to believe that restrictions on the multimediality and interactivity of e-learning is not possible without a negative impact on their quality. In this regard, relevant research should be conducted. The result of such a study would define the point at which the simplification of the multimedia objects in e-learning courses leads to an unacceptable level of quality. It would also be desirable to conduct studies comparing the perceived attractiveness of e-learning courses prepared on the basis of the same material but with three different approaches: a traditional approach, one based on RID and one which does not adapt a script as multimedia and interactive e-learning material. The possibility to carry out appropriate research is significantly impeded, since it requires financial resources for the development of alternative versions of an e-learning course.

VII. Conclusion

The paper presents a project management approach with integrated Rapid Instructional Design for e-learning course development and implementation. The verification of its effectiveness also validates the very concept of RID.

Study carried out during two e-learning projects confirmed both research hypotheses. Namely, adapting the concept of RID in approaches for the project management of e-learning course development supports the creation of e-learning courses at an acceptable level of quality while significantly reducing the time and cost of project realization compared with traditional approaches. In addition, the use of RID is a useful alternative in the development of e-learning courses, compared with preparing a simple e-learning programme based on static documents or expensive multimedia and interactive e-learning materials.

The validation results revealed the benefits of the approach developed by integrating the concept of RID, in that it develops e-learning courses at an acceptable quality, at a fraction of the cost and time compared with traditional approaches. The proposed solution enables organizations such as universities and training companies to prepare a wider e-learning programme. Therefore, in conjunction with confirmed two stated hypotheses, there is a positive answer to the question posed in the title of the article - Does the integration of the concept of Rapid Instructional Design in project management approaches support the efficient realization of e-learning projects?

The use of the developed approach during the CaseSimulator and CTF projects, combined with its validation results, showed the limitations of RID concept. RID-based approaches may be applicable to the preparation of e-learning courses of low or moderate complexity. It is highly problematic to adapt RID for projects where it is necessary to produce high quality e-learning packages in terms of multimediality and interactivity. In this case, the design and production processes should follow a more traditional approach.

References

[1] A. Akram, “Semi-Virtual Knowledge Engineering: Development of Semi-Virtual Knowledge Learning Process to Improve the Semi-Virtual Individual learning”, Communications of the IIMA: Vol. 10: Iss. 2, 2010.

[2] T. Bates, Technology, Open Learning and Distance learning, Routledge, 1995.

[3] S. Braxton, K. Bronico and K. Looms, Instructional design methodologies and techniques, The George Washington University, 1995.

[4] Catching the Future, http://catching.ug.edu.pl.

[5] Case Simulator, http://casesimulator.pl.

[6] E. T. Chen, “Successful E-Learning in Corporations”, Communications of the IIMA: Vol. 8: Iss. 2, 2008.

[7] R. Clark, Developing Technical Training: A Structured Approach for Developing Classroom and Computer-based Instructional Materials, Wiley, 2008.

[8] R. Clark, R. Mayer, e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning, Pfeiffer, 2007.

[9] J. Elen, Blocks on the Road to Instructional Design Prescriptions: a Methodology for I.D. – Research Exemplified, Leuven University Press, 1995.

[10] C. G. Gentry, Introduction to instructional development: Process and technique, Wadsworth Publishing Company, 1994.

[11] R. Hassel-Corbiel, Developing Technical Training Courses A technical Writer’s Guide to Instructional Design and Development, Learning Edge Publishing, 2001.

[12] B. Khan, Managing E-learning Strategies: Design, Delivery, Implementation and Evaluation, IGI Global, 2005.

[13] M. Kuciapski, “A design framework for instructional design information system”, Polish Journal of Environmental Studies,18(3B), 2009.

[14] M. Kuciapski, “Model for Project Management for Development and Implementation of E-Learning Courses”, Perspectives in Business Informatics Research, Proceedings of 9th International Conference, Springer, 2010.

[15] B. Locke, R. Perkins, K. Potter, J. Burton and G. K. Kreb, “Defining Quality in Distance learning: Examining National and International Standards for Online Learning”, Conference Proceedings 27th Annual Conference on Distance Teaching & Learning, Madison, 2011.

[16] G. M. Marković, B. Kliček and D. P. Vukovac, “The Effects of Multimedia Learning Materials Quality on Knowledge Acquisition”, Information Systems Development: Transforming Organisations and Society through Information Systems, 2014.

[17] M. McVay Lynch and J. Roecker, Project Managing E-Learning: A Handbook for Successful Design, Delivery and Management, Routledge, 2007.

[18] G. Piskurich, Rapid Instructional Design: Learning ID Fast and Right, Pfeiffer, 2006.

[19] G. Piskurich, Rapid Training Development: Developing Training Courses Fast and Right, Pfeiffer, 2009.

[20] D. Renner, S. Laumer and T. Weitzel, Effectiveness and Efficiency of Blended Learning – A Literature Review, IS in education, IS curriculum, education and teaching cases, AMCIS Proceedings, 2014.

[21] V. Ruhe and B. Zumbo, Evaluation in Distance learning and E-learning: The Unfolding Model, The Guildford Press, 2009.

[22] J. Sims, P. Powell and R. Vidgen, “Identifying E-Learning capabilities and competences”, UK Academy for Information Systems Conference Proceedings, 2013.

[23] S. Thiagarajan, “Rapid Instructional Design”. [Online]. Available from: http://www.thiagi.com/article-rid.html 2015.10.04.

[24] L. Zapf, “Building Scalable and Context-Dependable Repositories for Learning Objects using Open Source Components”, Learning Objects: Standards, Metadata, Repositories, and LCMS, Informing Science Press, 2007.

[25] W. Zhengui, “Identification and prioritisation of variables influencing the cost of learning content development”, Computers & Telecommunications Engineering, University of Wollongong, 2009.

On Behavioral Process Model Similarity Matching: A Centroid-based Approach

Michaela Baumann*, Michael Heinrich Baumann†, and Stefan Jablonski*
*Institute for Computer Science
†Institute for Mathematics
University of Bayreuth, Germany
Email: {michaela.baumann,michael.baumann,stefan.jablonski}@uni-bayreuth.de

***Abstract*—As business process models have a broad scope of applications, e.g., in science or in business administration, the problem of handling large amounts of process models arises. One helpful tool for dealing with this amount of models is to reduce it by using similarity measures in order to detect similar models that can be merged. A set of similar models may be replaced by one model. As a pure similarity of labels is often not enough to compare process models, other process perspectives are involved for calculating similarities. The current paper works on the process models' behavior, which is one such perspective. A problem that arises when comparing two models and that is covered in this paper is that one of a differing granularity of process steps. Due to this granularity problem M-to-N mappings are considered. The present paper provides a centroid-based and so easily computable method for calculating behavioral similarity values for process models, which is constructed for M-to-N mappings, and a short evaluation of it.**

Keywords–Business process model; Behavioral process model similarity; M:N-Matching

I. INTRODUCTION

Not only for documentation purposes, business process models have been established in a large amount of organizations. They also serve as supportal means for communication, for training employees, and redesigning actual workflows [1]. These widely spread applications lead to vast process model repositories in enterprises, that have to be managed somehow [2]. One of these management purposes is to find similar models in order to reduce the tremendous amount of repository elements by detecting and merging similar models. Similar models can emerge when the same process is modeled multiple times, either for different end user groups or in different variations for the same user group. The authors of [3] worked out a total of nine categories for application fields of similarity measures, amongst them process merging, facilitating reuse of models [4], and service discovery. Also, process model matching can be used in the fields of compliance and conformance checking, the latter especially in terms of (process) log data, which can be seen as a number of sequential process models. But usually, the models are developed by different persons and thus have different levels of granularity, which means, that process steps in different models are modeled with a different fineness [5]. Especially for human tasks, it is not prescribed how fine-grained the particular steps have to be, and the detail level strongly depends on the purpose the model has to fulfill, on the attributes of the executing agents, or simply on the modeler's preferences. Furthermore, the terminology, i.e., the way of defining names, labels, etc. varies from model to model, and hence a comparison of these models only using their labels is challenging [6]. These two issues often lead to the fact that actually very similar or even equal models are not recognized as such. Because of this and due to the wide variety of modeling languages and notations, like Event-driven Process Chains (EPCs), Petri Nets, Unified Modeling Language (UML) Activity Diagrams, Workflow Nets, the Business Process Model and Notation (BPMN), or the Business Process Execution Language (BPEL), perfect matches, i.e., a true/false answer to the question if two models are the same, cannot be expected. Instead, a degree of similarity, a value between 0 and 1 where 0 means completely different and 1 is an indication for (virtually) identical models, depending on the definition of the respective similarity measure, is desired.

These measures can be defined on different disjoint aspects of process models: on node information, on process structure, and on execution semantics [2]. Node information is attached to each process model element, especially activities, and can again be split up into the description of process model elements, assigned roles or agents, ingoing and outcoming data objects, and operational means. Process structure refers to graph structure when taking a process model as a graph, and execution semantics refers to the question, how, i.e., in which order and under which circumstances (parallel, inclusive, exclusive, loop, etc.), process model elements may be executed. A behavioral similarity usually relies on the execution semantics of a process model. In order to take into account all of this information about process models and to allow for a wide range of modeling notations, we define a process model as instanced in Definition 1. In principle, each process model, whatever modeling notation is used, is a graph consisting of bubbles and directed arcs. Bubbles are elements like activities, events, and gateways, connected through arcs. Execution order is thus more or less prescribed, and human influence, i.e., decisions, are involved through (exclusive or inclusive) gateways. Note that this holds for imperative process models. Declarative process models, like the Case Management Model and Notation (CMMN) [7], take a different approach and allow for a greater human influence.

Definition 1 (Process Model)**.** *A process model is a tuple* $G = (N, E, \lambda, \delta)$ *where* $N = A \cup \{start, end\} \cup S_{AND} \cup S_{XOR} \cup S_{OR}$ *is the finite, non-empty set consisting of all model elements, and* $E \subseteq N^2$ *is the set of all directed edges connecting the elements of* N*. Function* λ *assigns a description, a data set, organizational, and operational resources to each of the tasks in* A*. Function* δ *assigns constraint descriptions to some edges.*

Set S_{AND} is the set of all (split and merge) parallel gateways. Sets S_{XOR} and S_{OR} are the sets of all (split and merge) exclusive or inclusive gateways, respectively. $start$ denotes the start event of the process and end the end event. Every process has exactly one start and one end event. The activity tasks are summarized in set A. Functions λ and δ are mentioned for the sake of completeness, but are not discussed further, as they are not of importance for the behavior of a process model. Task description, used data, assigned agents, assigned tools and behavior can be treated separately when analyzing process models, as these five perspectives are completely orthogonal to each other [8]. Similarity of descriptions can be determined via string-edit operations, see for example [9], whereas data, organizational and operational similarity can be calculated with set-based methods, like the Jaccard coefficient [10] [11]. A more detailed definition of multi-perspective process models can be found in [12]. The elements of set N are sometimes also called nodes.

Definition 1 allows for many kinds of process models, even if they do not provide information about all process perspectives. For instance, if the non-human resources, that is the operational perspective, is not given in the model, the corresponding co-domain of function λ is left empty. Or if inclusive gateways are not included, then it is set $S_{OR} = \emptyset$. Human influence on the behavior is covered by exclusive (XOR) and inclusive (OR) gateways and the agents' decisions during the execution. At design level, however, this influence, i.e., the decision at run time, does not affect the model behavior. In imperative process models, behavior is strongly restricted.

The focus of the work at hand lies on the behavioral aspect of process models, i.e., on control flow and how two models can be compared with respect to this aspect. During the matching process – this is what we call the process of finding a similarity value between two models – the tasks of two process models are not compared one-to-one (single task compared to single task), but they will be grouped into sets to encounter the problem of differing granularity. In many cases, one-to-one mappings are not able to represent the correct correpondences. For example, when one activity in the first process model is split up into three process steps in the second model (imagine a manager's and a technician's view on a certain process), a one-to-one mapping would not provide a satisfying result. After having established the task sets, centroids, i.e., average positions (see Definition 4), average repeatability, and average optionality are calculated to determine behavioral similarity. As far as the authors know, this distinction of behavior into the three dimensions position, repeatability, and optionality has not yet been done explicitely in previous work.

In [13], process model elements are classified into, among other things, alternative or loop fragments, that resemble optional and repeatable elements. Furthermore, these centroids will be able to punish sets of activities that are widely spread over the whole process model or that have strongly differing manner. See Figure 1 for an example of two process models with schematical positional centroids. The mapped task sets are indicated with different fillings. The resulting behavioral similarity value can then be combined with other similarity values, e.g., description similarity or data similarity, to get a better matching score that is more independent of local errors, i.e., that is more robust against errors in certain process model aspects [10]. To put it together, the method presented in this paper provides two main results: A normalized similarity value for two process models based on their behavior and a mapping that indicates the resembling parts of them, which will be called M-to-N mapping. The mapping is needed to compute the behavioral similarity value. This approach is also known in related work, e.g., in [2] and [14] using 1-to-1 mappings. The advantage of such a method compared to a pure similarity calculation without presupposing a mapping is that the correspondences are provided in the same step and do not have to be detected in a separate step afterwards. The M-to-N mapping, also used in [11] for organizational and operational similarity, is defined in Definition 2.

Definition 2 (M-to-N mapping)**.** *Let $G_i = (N_i, E_i, \lambda_i, \delta_i)$, $i = 1, 2$ be two process models, with $A_i \subset N_i$ being the set of activities or tasks of each process model and $P_i \subset \mathcal{P}(A_i) \ni \emptyset$ a complete and disjoint partition of A_i, $i = 1, 2$. A mapping between G_1 and G_2 is defined as a bijective function $M : P_1 \to P_2$. In particular, $\emptyset \mapsto p_2$ and $p_1 \mapsto \emptyset$ means, that p_2 and p_1 are deleted, respectively, where $p_1 \in P_1$, $p_2 \in P_2$, and $\neg(\emptyset \mapsto \emptyset)$.*

As Definition 2 shows, sets of activities are mapped rather than single tasks, which induces the term M-to-N mapping. These sets of activities are achieved by establishing a partition of set A. In Figure 1, the tasks of the left model are partitioned into four sets (one of them the empty set), as well as the tasks of the right model. Tasks are indicated through rectangles with rounded corners, the diamonds represent gateways. Diamonds filled with "x" are exclusive, filled with "+" are parallel, filled with a small ring inclusive. In Figure 1, the meaning of the gateways is not of importance. Start and end event are denoted by circles. The mapping consists of four elements, a 2-to-3 (dotted) and a 2-to-1 (striped) mapping element, as well as a 1-to-0 (white) and a 0-to-1 (gray) element (the deleted nodes). In cases of process models strongly differing in granularity, a comparison explicitly applying a M-to-N mapping may provide better results than methods presented in most of the previous work. Furthermore, no complex calculations are needed for the centroid-based similarity presented in Section III. Regarding the draft version of this paper [12], requirements for compared process models, like block-structure, have been relaxed.

The remainder of this paper is organized as follows: The next section gives a rough overview of existing similarity measures and process model matching methods. Section III then introduces the behavioral similarity measure in its three dimensions step by step. An extension for penalized similarity measures is given, too. Thereafter, in Section IV, a short evaluation is performed. Section V revises the paper and gives ideas for future work.

II. Background and Related Work

In the literature, many techniques and methods for calculating the similarity, or, on contrast, the distance of process models, are presented. The authors of [3] provide a comparing overview of some of these techniques. Other collections and comparisons of several matching techniques can be found in [15] and [16]. One way of measuring the similarity between a pair of process models is to first define a mapping between

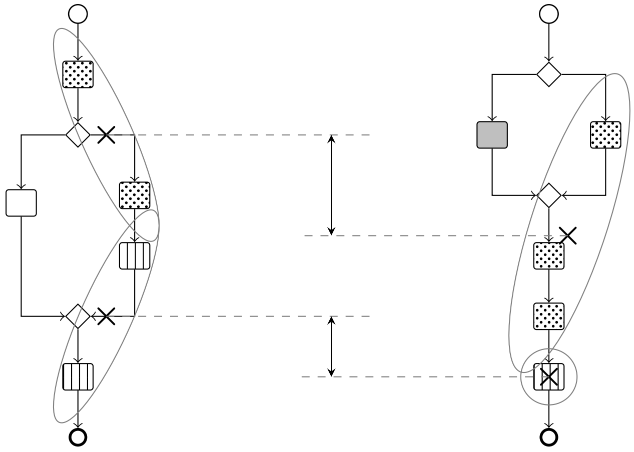


Figure 1: Schematical representation of the comparison of (positional) centroids for mapped sets of process tasks.

these two models. This mapping can either assign one node of the first model to one node of the second model, which often leads to a partial injective function [14], or map a set of nodes of one model to a set of nodes of the second model [10]. Thereby, we will refer to the latter defined mapping as M-to-N mapping, which is defined according to Definition 2 and is just a generalization of the former 1-to-1 mapping definition through an extension to powersets. As the authors of [9] suggest, for many scenarios, e.g., when processes have been developed independent of each other, a M-to-N mapping is preferred to a simple 1-to-1 mapping. M-to-N mappings are capable to overcome problems of granularity levels, which is one of the future tasks stated in [17]. In [6], a method for establishing N-to-1 mappings is presented, but not extended to a M-to-N mapping due to the applied matching techniques.

A. Label-based and structural similarity values

After having established a mapping between the elements of process models, similarity values between these elements can be computed. Depending on the given models, various information is used for this computation step. A similarity value based on the activity labels of a process model is, e.g., presented in [14], [18], and [19]. It makes use of the so-called string-edit (Levenshtein) distance and other string-modifying techniques like stemming [14] or replacing certain words through synonyms [20]. This similarity is often referred to as syntactic, semantic, or linguistic similarity. Another information that can be used for comparing process models is information about authorized agents and assigned input and output data of each activity, which is available, for example, in BPMN process models [21]. In Definition 1, this information can be found in function λ, that maps each node to a four-tuple consisting of node description, authorized agents, tools to be used, and consumed and produced data. This additional information can be analyzed lexically [6] or when applying M-to-N mappings through set-based methods performed on the subjects' or objects' identifiers, as it is done in [10] and [11]. Another important aspect of process models is the arrangement of the model elements. Basically, this arrangement can be categorized into two different similarity metrics: structural/contextual similarity and behavioral similarity [9] [22]. Structural and contextual similarity is, however, not using a preceding mapping, so we leave these kinds of measures out as we want to use a similarity measure showing the model equivalences at the same time. Behavioral matching techniques are based on the execution semantics of a process model [14], which means, that, e.g., parallelism or exclusiveness of model elements as well as their possible execution order is respected.

B. Various Definitions of Behavioral Similarity

Different approaches for measuring behavioral similarity are developed in literature. In [10], a computing method for M-to-N mappings is suggested that makes use of partial order relations of the activity elements, but is limited to serialized process models without any gateways, which is a strong limitation for most models. Behavioral profiles, a set of valid relations (strict, exclusive, interleaving) between every two process model elements, are introduced in [22] and [23] to define different behavioral similarity values. This approach is, however, applicable if the process models are mapped 1-to-1 and hardly transferable to M-to-N mappings. Another common method is to look at the traces of the process models to be compared [2]. Even if there is only a finite set of traces in loop-free process models, the problem of computing the trace-based behavior of a model is NP-hard [9]. An explicit discussion of trace-based methods is presented in [3]. Regarding partial traces is a variant of this trace-based approach and discussed in [24]. To overcome the computational complexity of traces, an approximation via casual footprints can be performed [2]. Casual footprints define the so-called look-back and look-ahead links of single process model elements [9] [20] and not for sets of tasks. Furthermore, this similarity value takes sequential, parallel, or exclusive behavior of the model elements, which is important information about a process model's behavior, only insufficiently into account [9]. A further approach of determining similarity between two process models is given in [25], where process models are compared with respect to some typical behavior that is gained from process event logs. However, as event logs or typical reference models are not always given, the applicability is restricted to particular cases. The aim of the paper is to derive a behavioral similarity function whose calculation is not too difficult, in contrast to the calculation of casual footprints [3], and that is at the same time suitable for M-to-N mappings. All in all, an approach like ours makes use of already existing concepts like embeddability into graph-edit similarity, but is adjusted to situations where previous approaches are not able to detect similarity or to take into account all given information.

Another work dealing with comparison of workflows, workflow systems, and the expressiveness of workflow languages is [26]. It is very broad, but uses simulation and not the models itself for comparison. When using simulation, there is the difficulty of finding significant samples and the fact, that all results are statistical, i.e., hold under a certain significance level. It also introduces a lot of notion and transformation methods to tree and automaton representations for the models. Furthermore, the different process perspectives are not considered separately and concrete correspondences are not worked out explicitely. The authors of [27] also worked out a method to determine behavioral similarity based on arbitrary alignments (thus, also for M-to-N mappings), where overlapping of the mapped node sets is allowed, which is a great advantage of this work. In return, they only allow for acyclic, i.e., loop-free, process models based on causal nets. Repeatability is therefore not considered.

No matter what kind of mapping is established and which kind of information (label, behavior, etc.) is used to calculate similarity, the further progress is always the same. After having computed the similarity value for one particular mapping, this value is maximized over all possible mappings to get the best correspondences between the two models. I.e., $Sim(G_1, G_2) = \max_M Sim_M(G_1, G_2)$ gives the similarity value for the two models G_1 and G_2 where function Sim_M calculates their similarity induced by mapping M [10] [14]. For this optimization step, greedy or A* algorithms working on task set A in the case of 1-to-1 mappings or on the powerset of A in the case of M-to-N mappings can be used [18].

III. The Centroid-based Behavioral Similarity Measure

For our work, we rely on process models given according to Definition 1, that even may contain loops. Models with loops particularly allow for control loops. The similarity calculation further assumes a M-to-N mapping between the two models that shall be compared and makes use of the centroids of the task sets. In particular, a M-to-N mapping M is given according to Definition 2, i.e., a partition P_1 of activities A_1 of the first process model G_1 is mapped bijectively to a partition P_2 of activities A_2 of the second process model G_2, and every element of a partition $p \in P_i$ is a set of activities of the underlying process model, i.e., $p \subseteq A_i$.
The targeted behavioral similarity measure considers the order of nodes given by the control flow, but also takes into account mandatory and optional activities as well as repeatable ones, which we call the three dimensions of behavior as already mentioned in the introduction. A penalty score is added to neglect sets of heterogenous tasks, e.g., widely spread sets of tasks.

A. Positional Similarity

The first behavioral dimension reflects the location of nodes in a process model. This location is specified as a relative position to obtain comparability, i.e., the position of a node is a number in $[0, 1]$, where a value close to zero indicates a position at the beginning of the model and a value close to one a position at the end. In particular, the position of a node is given through the length of the shortest chain (sequence of consecutive directed edges) from the $start$ event to the node, divided by the length of the shortest chain going from $start$ to end while passing the node. This is specified in Definition 3. Function $m(\cdot, \cdot)$ in Definition 3 gives the length of the shortest chain from one node to another. By using these minimal chains the problem of infinitely long chains resulting from loops is avoided. The position of a set of nodes, i.e., the positional centroid, given in Definition 4, is then computed by simply taking the arithmetic average of the single positions, i.e., summing up the single positions and dividing through the number of elements in the set. This again results in a value in $[0, 1]$. Note that the definitions basically apply for all nodes, like events, tasks and gateways, but we will later on use only the activity tasks' positions, as only the tasks are mapped by a M-to-N mapping according to Definition 2.

Definition 3 (Node position)**.** *The position $\pi(n)$ of node $n \in N$ is $\pi(n) = \frac{m(start,n)}{m(start,n)+m(n,end)}$.*

Positions that are fixed in all process models we consider are $\pi(start) = 0$ and $\pi(end) = 1$ as $m(start, start) = m(end, end) = 0$. In the following definition, P denotes a partition of a model whose elements p are sets of nodes, i.e., $p \subseteq N$.

Definition 4 (Centroid of a set of nodes)**.** *The centroid $\pi(p)$ of $p \in P$ is given through $\pi(\emptyset) = \texttt{NULL}$ and*

$$\pi(p) = \tfrac{1}{|p|} \sum\nolimits_{n \in p} \pi(n), \; p \neq \emptyset. \tag{1}$$

All occuring $NULL$-values are ignored in the further calculations in this paper, but lower the overall similarity when combining the behavioral similarity with other kinds of similarity like label- or resource-based similarity. The $NULL$ values occur if nodes are not mapped, like it is the case in Figure 1 for the white and the gray task. The behavioral similarity of two models, represented by their partitions P_1 and P_2, then combines the differences of the centroids of the mapped sets of tasks again as an arithmetic mean.

Definition 5 (Behavioral similarity 1)**.** *For two partitions P_1, P_2 of process models G_1, G_2 induced by a mapping M, the first dimension of behavioral similarity, the position-based similarity, is given through*

$$VSim^{\pi}_M(P_1, P_2) = \tfrac{1}{|P_1|} \sum\nolimits_{p \in P_1} (1 - |\pi(p) - \pi(M(p))|) \,. \tag{2}$$

Figure 1 shows two centroid differences: $|\pi(p_{dotted}) - \pi(M(p_{dotted}))|$ and $|\pi(p_{striped}) - \pi(M(p_{striped}))|$. Low differences, i.e., similar positions, lead to high similarity values due to the modification $1 - |\cdot|$ in formula (2). The formula can also be formulated with the models themselves via $VSim^{\pi}_M(G_1, G_2) := VSim^{\pi}_M(P_1, P_2)$, as mapping M applied on the models induces the partitions.

B. Repeatability and Optionality

Besides the position value π, we can also assign a repeatability value ϱ and an optionality value o (*omikron*) to each node. These additional dimensions of the behavior of process models display the execution with regard to the different gateway types. The approach for these two is similar to that of Section III-A. First, a repeatability/optionality value is defined for single nodes. Then, a repeatability/optionality value for a set of nodes is established through the arithmetic average of the single values. Finally, these values are combined for the partitions induced by the mapping.

Definition 6 (Node repeatability)**.** *The repeatability $\varrho(n)$ of node $n \in N$ is $\varrho(n) = 1$ if n can be executed more than once in one process instance and 0 otherwise.*

The repeatability value provides information if a node can be executed more than once in one process instance, i.e., if it is involved in a XOR-loop. In BPMN, it is possible to mark activities as loop tasks which are also treated as repeatable nodes. Another property of nodes is their optionality, i.e., if a node has to be executed in one process instance or if the process can finish without having executed it. Optionality can be given if XOR- or OR-gateways appear.

Definition 7 (Node optionality)**.** *The optionality $o(n)$ of node $n \in N$ is $o(n) = 1$ if n does not have to be executed to finish an instance of the process successfully, and 0 otherwise.*

Both repeatability and optionality values are boolean. As we do not assume any process log information about executed instances as e.g. shown in [25], there is no statement if an optional node is more or less likely to be executed or how often a repeatable node is executed in average. For future work, one can think of also assigning optionality/repeatability values $\in (0, 1)$, e.g., by using execution probabilities or relative frequencies obtained from process execution logs. Analog to Definition 4, repeatability and optionality is extended to sets of nodes as shown in the following definition.

Definition 8 (Repeatability and optionality of node sets)**.** *For $p \in P$, P a partition of G (i.e., $p \subseteq A$), repeatability $\varrho(p)$ and optionality $o(p)$ of a node set p is given through equation (1) by replacing π through ϱ or o, respectively.*

With this, behavioral similarity for the two remaining behavior dimensions can be formulated.

Definition 9 (Behavioral similarity 2 and 3)**.** *For two partitions P_1, P_2 of G_1, G_2 induced by a mapping M, the behavior similarities based on repeatability and optionality is given through equation (2) by replacing π through ϱ or o, respectively.*

C. Penalty Functions

The positional centroids of a task set consisting of "the first" and "the last" task and of a task set consisting of exactly one task in the middle of the model would be the same, when calculated according to formula (2), namely 0.5. But it is quite obvious, that these two sets of nodes are unlikely to match together (regarding their behavior). This is why we introduce penalty terms for every dimension of behavioral similarity, that lower the similarity value if one or both partition elements p and $M(p)$ are heterogenous activity sets. Especially for favouring homogeneity concerning repeatability and optionality in node sets, penalty functions are important. These functions depend on the underlying mapping M and are denoted with $pen^{\pi}_{M}, pen^{\varrho}_{M}, pen^{o}_{M} \geq 0$. They have to be computed for each partition separately. The resulting penalized similarity is of the form $penVSim^{\xi}_{M}(P_1, P_2) = \Big(VSim^{\xi}_{M}(P_1, P_2) - pen^{\xi}_{M}(P_1) - pen^{\xi}_{M}(P_2)\Big)^{+}$, where $\xi \in \{\pi, \varrho, o\}$ and P_1 and P_2 are the partitions induced by M on the two process models G_1 and G_2. We set $penVSim^{\xi}_{M}(G_1, G_2) := penVSim^{\xi}_{M}(P_1, P_2)$.

As $VSim \in [0, 1]$ it is reasonable to demand for penalty functions $pen \in [0, 0.5]$. A function that meets this requirement and that somehow measures the spread of a set of objects is the variance, in this case the sample variance, that uses the centroids as (sample) means. Therefore, if we apply the unbiased sample variance, we get $pen^{\xi}_{M}(p) = \frac{1}{|p|-1}\sum_{a \in p}(\xi(a) - \xi(p))^2$ with $\xi \in \{\pi, \varrho, o\}$ as penalty value for one partition element $p \in P$ with $|p| \geq 2$. For $|p| = 1$ the penalty value is 0 and for $p = \emptyset$ it is not available, i.e., set to $NULL$. The penalty value for a whole partition P is computed as the average over the single penalty values $pen^{\xi}_{M}(P) = \frac{1}{|P|}\sum_{p \in P} pen^{\xi}_{M}(p)$.

D. (Penalized) Behavioral Similarity

To get one value for behavioral similarity, one has to combine the three dimensions of behavior and their corresponding similarity values $VSim^{\pi}$, $VSim^{\varrho}$, and $VSim^{o}$ or, analog, the penalized similarity values $penVSim^{\pi}$, $penVSim^{\varrho}$, and $penVSim^{o}$. This combination can take place with help of a weighted sum of the three values, where the weights can be chosen according to one's own impression of suitability or, which would be worth futher studies, according to statistical findings including model training and parameter estimation, e.g., maximum likelihood methods. With non-negative weights w^{π}, w^{ϱ}, and w^{o} with $w^{\pi}+w^{\varrho}+w^{o} = 1$ the weighted sum, i.e., the behavioral similarity value for two process models G_1 and G_2 under mapping M, is of the form $VSim_{M}(G_1, G_2) := \sum_{\xi \in \{\pi, \varrho, o\}} \omega^{\xi} VSim^{\xi}_{M}(G_1, G_2)$.

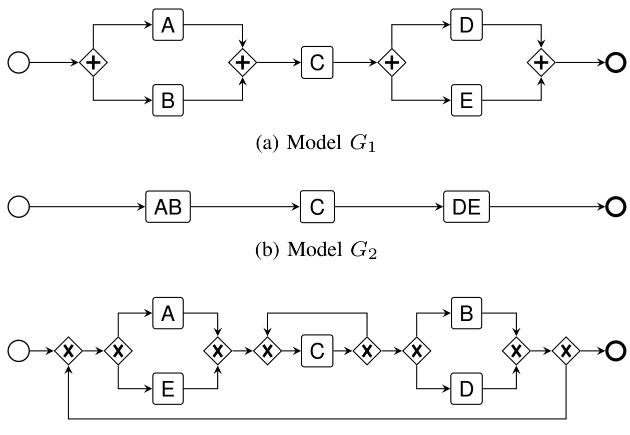
For the penalized behavioral similarity $penVSim_{M}(G_1, G_2)$, the similarity values for the three behavioral dimensions are replaced by their respective penalized similarity values. Both $VSim$ and $penVSim$ always take values between 0 and 1 where 0 means no similarity and 1 full similarity. The (penalized) behavioral similarity can then again be used for calculating the similarity value including other process model perspectives [11].

IV. Validation

A comparison of three methods to measure behavioral similarity is done in this section. Therefore, three process models G_1, G_2, and G_3, shown in Figure 2, are considered. Models G_1 and G_2 describe the same process, but were modeled by different persons. Model G_3 describes a different process including similar tasks, but with a differing order. In G_3, not all tasks have to be executed and some may be executed several times. Models G_1 and G_2 always have activities A to E executed exactly once. The original label descriptions have been removed and substituted by letters A to E to provide better readability, as the focus lies only on the models' behavior. Information about agents, non-human resources, and data is not shown in the models, either.

For calculating the similarity between Models G_1 and G_2, the *p*artial *i*njective 1-to-1 mapping M^{pi}_{1} is established with $\{(A, AB), (C, C), (D, DE)\} = M^{pi}_{1}$. Tasks B and E from G_1 are not mapped, but results would not differ if B and E instead of A and D would have been mapped. The *b*ijective M-to-N mapping M^{b}_{1} according to Definition 2 is established with $\{(\{A, B\}, \{AB\}),\ (\{C\}, \{C\}),\ (\{D, E\}, \{DE\})\} = M^{b}_{1}$. These mappings provide the highest similarity, respectively, when taking into account the activities' descriptions (using string-edit distance). The mappings for the comparison of G_1 and G_3 are the identity functions, namely $M^{pi}_{2} = \{(\cdot, \cdot) \mid \cdot \in \{A, B, C, D, E\}\}$ and $M^{b}_{2} = \{(\{\cdot\}, \{\cdot\}) \mid \cdot \in \{A, B, C, D, E\}\}$.

For evaluation, three behavioral similarity values are computed for every comparison. One with help of casual footprints (CF) [9], one with smallest casual footprints (smallest CF) as suggested in [3], Section 6.3, IV *discussion*, and one with the



(a) Model G_1

(b) Model G_2

(c) Model G_3

Figure 2: Behavioral similarity values are computed for models G_1–G_2 and G_1–G_3

penalized centroid-based approach as introduced in the present work. The centroid-based method uses M_1^b and M_2^b and the positions (or rather the positional distances) of the respective sets of nodes induced by these mappings. The smallest CF is a modified casual footprint approach with a slightly different, especially simplified, definition. The results of the computation are listed in Table I. The numbers in brackets count the effort of determining the respective similarity values. In particular, the number of computed intermediate values are specified. As the calculation of casual footprints needs another underlying similarity value, called correlation, between the compared nodes, we chose a label-based similarity. For simplicity, it was set $Sim(A, AB) = Sim(D, DE) = 0.5$ and $Sim(\cdot, \cdot) = 1 \; \forall \cdot \in \{A, B, C, D, E\}$.

Obviously, all three methods state that models G_1 and G_2 are more similar than models G_1 and G_3, which is as desired and also assessed by several modeling experts. But differences between the assigned similarity values are substantial. The centroid-based approach states full behavioral similarity between models G_1 and G_2, which could be discussed if this result fits reality, because the first model's parallel gateways seem to be ignored. The casual footprint method instead says similarity is only about 80%, although, as stated above, when finished, all activities A to E are executed exactly once in both models. In contrast, the casual footprint method assigns a similarity of about 64% to G_1 and G_3, although these models describe completely different processes concerning their behavior. The centroid-based approach assigns a relatively low similarity value of about 33%. For both comparisons, the smallest casual footprint approach gives values in between the two other methods. Another even more remarkable difference gets apparent when considering the number of calculated intermediate values (not elementary arithmetical operations). They are shown for all three methods and both comparisons in brackets in Table I. It is apparent, that between the common casual footprint method and the smallest casual footprint approach there is a huge difference in the number of calculated intermediate values, even if the resulting similarity values do not differ that much. For the casual footprint method, the number of intermediate values rises exponentially with the number of model nodes. For the smallest casual footprint approach, this number is only increasing quadratic, which was one of the reasons for the authors of [3] to introduce it. For the centroid-based approach, the number of calculated intermediate values rises linearly with the number of model activity nodes, so the effort is even less, which is a strong point for this method.

TABLE I: Similarity values and number of computed intermediate values (IM).

$Sim.\ (\#IM)$	CF	smallest CF	centroid-based
$Sim(G_1, G_2)$	0.799 (294)	0.885 (90)	1.000 (30)
$Sim(G_1, G_3)$	0.640 (414)	0.632 (108)	0.333 (30)

It should be pointed out again that the centroid-based similarity value gives information about only one aspect of the compared process models. Information about labels, data, and resources is not used for calculating this value. Similarity values concerning these aspects can be calculated separately and then be combined altogether. Instead, the casual footprint method needs a similarity value assigned to each pair of activities which is element of the underlying mapping. Thus, the casual footprint method does not completely separate the different process perspectives orthogonally from each other. However, the results of similarity calculations using the centroid-based method only make sense when combined with other perspective similarities. Otherwise, in the majority of the cases, it would lead to a mapping where first node is mapped to first node, second node to second node, and so on. The behavioral similarity is more like an endorsement of the mappings and similarities obtained for the other perspectives, that also consider the content of the model elements (labels, agents, etc.), which is not the case for the behavioral aspect. This is why a more comprehensive evaluation of the method can only take place together with the other process perspectives, which has to be done nonetheless to fully evaluate the centroid-based approach. As possible evaluation setting, a similar setup as in [9] would probably be a good choice. The draft version for this paper [12] contains a second example for application of the centroid-based similarity measure.

V. Conclusion and Future Work

As shown in Section II, there already exists a variety of techniques for calculating the behavioral similarity of arbitrary process models. The methods presented in the work at hand should not be seen as strictly better, but should rather help in computing *behavioral similarity for M-to-N mappings*, for which behavioral similarity measures applicable on general process models did not exist. A big advantage of the *centroid-based approach* is that average values can *easily be computed*, unlike trace-based and casual footprint methods, and thus, the presented method is suitable even for large practical applications. Furthermore, the idea of splitting process models into several perspectives like label description, data objects, etc., as already pursued in multiple similarity matching papers, is continued in this work by dividing model behavior into the three dimensions *(relative) position, repeatability, and optionality*, which allows for adjusting the weighting of these three dimensions according to the user's needs. The separation of behavior is not done in related work, as far as the authors know. *Penalty terms* are a common means in the field of measure construction, although in the context of process model similarity measuring they have not been used so far. However, the centroid-based similarity measure is not very informative when used on its own. It has to be combined with similarity values for other process perspectives, but this is also the case for, e.g., casual footprints.

Some approaches for future work are already stated in the main part of the paper. Concerning execution specific features of process models, e.g., the optionality value of a node, it is conceivable to use process log information to improve similarity values. Additionally, parameters, weights, and maybe even formulae might be improved by applying machine learning methods on already matched process models. Another big topic for future work would be to implement M-to-N matching methods for all process model perspectives and to run a detailed evaluation that combines all similarity values for the different perspectives. Furthermore, it could be checked if the centroid-based approach provides a real metric, i.e., if it fulfills the corresponding conditions of symmetry, non-negativity, identity, and the triangle inequality. Or, if this is not the case, can it be (easily) adjusted to achieve these properties, as with such metrics it is possible to search huge repositories even faster for similar elements, e.g., with metric trees [22].

ACKNOWLEDGEMENT

The work of Michael Heinrich Baumann is supported by a scholarship of "Hanns-Seidel-Stiftung (HSS)" which is funded by "Bundesministerium für Bildung und Forschung (BMBF)".

REFERENCES

[1] M. Dumas, M. La Rosa, J. Mendling, and H. A. Reijers, *Fundamentals of Business Process Management*. Springer, 2013.

[2] M. Dumas, L. García-Bañuelos, and R. M. Dijkman, "Similarity search of business process models," *IEEE Data Eng. Bull.*, vol. 32, no. 3, pp. 23–28, 2009.

[3] M. Becker and R. Laue, "A comparative survey of business process similarity measures," *Computers in Industry*, vol. 63, no. 2, pp. 148–67, 2012.

[4] F. Pittke, H. Leopold, J. Mendling, and G. Tamm, "Enabling reuse of process models through the detection of similar process parts," in *Business Process Management Workshops*, ser. LNBIP, M. La Rosa and P. Soffer, Eds. Springer Berlin Heidelberg, 2013, vol. 132, pp. 586–597.

[5] B. Curtis, M. I. Kellner, and J. Over, "Process modeling," *Commun. ACM*, vol. 35, no. 9, pp. 75–90, 1992.

[6] M. Weidlich, R. Dijkman, and J. Mendling, "The icop framework: Identification of correspondences between process models," in *Advanced Information Systems Engineering*, ser. LNCS, B. Pernici, Ed. Springer Berlin Heidelberg, 2010, vol. 6051, pp. 483–498.

[7] Object Management Group, "Case management model and notation version 1.0," 2014. [Online]. Available: http://www.omg.org/spec/CMMN/1.0/PDF/ [accessed: 2015-07-19]

[8] S. Jablonski and C. Bussler, *Workflow management: modeling concepts, architecture and implementation*. International Thomson Computer Press, 1996.

[9] R. Dijkman, M. Dumas, B. van Dongen, R. Käärik, and J. Mendling, "Similarity of business process models: Metrics and evaluation," *Information Systems*, vol. 36, no. 2, pp. 498 – 516, 2011.

[10] M. H. Baumann, M. Baumann, S. Schönig, and S. Jablonski, "Towards multi-perspective process model similarity matching," in *Enterprise and Organizational Modeling and Simulation*, ser. LNBIP, J. Barjis and R. Pergl, Eds. Springer Berlin Heidelberg, 2014, vol. 191, pp. 21–37.

[11] M. Baumann, M. H. Baumann, S. Schönig, and S. Jablonski, "Resource-aware process model similarity matching," 2014, in press (RMSOC).

[12] M. Baumann, M. H. Baumann, and S. Jablonski, "On behavioral process model similarity matching: A centroid-based approach," 2015, preprint. [Online]. Available: https://epub.uni-bayreuth.de/id/eprint/2051 [accessed 2015-07-18]

[13] C. Gerth, M. Luckey, J. Küster, and G. Engels, "Detection of semantically equivalent fragments for business process model change management," in *International Conference on Services Computing (SCC)*. IEEE, 2010, pp. 57–64.

[14] R. Dijkman, M. Dumas, L. García-Bañuelos, and R. Käärik, "Aligning business process models," in *International Enterprise Distributed Object Computing Conference*. IEEE, 2009, pp. 45–53.

[15] U. Cayoglu *et al.*, "The process model matching contest 2013," in *4th International Workshop on Process Model Collections: Management and Reuse, PMC-MR*, 2013.

[16] J. Starlinger, B. Brancotte, S. Cohen-Boulakia, and U. Leser, "Similarity search for scientific workflows," *Proc. VLDB Endow.*, vol. 7, no. 12, pp. 1143–1154, 2014.

[17] R. M. Dijkman *et al.*, "A short survey on process model similarity," in *Seminal Contributions to Information Systems Engineering*, J. Bubenko, J. Krogstie, O. Pastor, B. Pernici, C. Rolland, and A. Sølvberg, Eds. Springer Berlin Heidelberg, 2013, pp. 421–427.

[18] R. Dijkman, M. Dumas, and L. García-Bañuelos, "Graph matching algorithms for business process model similarity search," in *Business Process Management*, ser. LNCS, U. Dayal, J. Eder, J. Koehler, and H. A. Reijers, Eds. Springer Berlin Heidelberg, 2009, vol. 5701, pp. 48–63.

[19] C. Klinkmüller, I. Weber, J. Mendling, H. Leopold, and A. Ludwig, "Increasing recall of process model matching by improved activity label matching," in *Business Process Management*, ser. LNCS, F. Daniel, J. Wang, and B. Weber, Eds. Springer Berlin Heidelberg, 2013, vol. 8094, pp. 211–218.

[20] B. van Dongen, R. Dijkman, and J. Mendling, "Measuring similarity between business process models," in *Advanced Information Systems Engineering*, ser. LNCS, Z. Bellahsène and M. Léonard, Eds. Springer Berlin Heidelberg, 2008, vol. 5074, pp. 450–464.

[21] BPM-Offensive Berlin, "BPMNPoster," 2009. [Online]. Available: http://www.bpmb.de/index.php/BPMNPoster [accessed: 2015-07-18]

[22] M. Kunze and M. Weske, "Metric trees for efficient similarity search in large process model repositories," in *Business Process Management Workshops*, ser. LNBIP, M. zur Muehlen and J. Su, Eds. Springer Berlin Heidelberg, 2011, vol. 66, pp. 535–546.

[23] M. Kunze, M. Weidlich, and M. Weske, "m3–a behavioral similarity metric for business processes," in *ZEUS*, ser. CEUR-WS, 2011, pp. 89–95.

[24] A. Wombacher, "Evaluation of technical measures for workflow similarity based on a pilot study," in *On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE*, ser. LNCS, R. Meersman and Z. Tari, Eds. Springer Berlin Heidelberg, 2006, vol. 4275, pp. 255–272.

[25] W. van der Aalst, A. de Medeiros, and A. Weijters, "Process equivalence: Comparing two process models based on observed behavior," in *Business Process Management*, ser. LNCS, S. Dustdar, J. Fiadeiro, and A. Sheth, Eds. Springer Berlin Heidelberg, 2006, vol. 4102, pp. 129–144.

[26] S. Abiteboul, P. Bourhis, and V. Vianu, "Comparing workflow specification languages: A matter of views," *ACM Trans. Database Syst.*, vol. 37, no. 2, pp. 10:1–10:59, 2012.

[27] A. Polyvyanyy, M. Weidlich, and M. Weske, "Isotactics as a foundation for alignment and abstraction of behavioral models," in *Business Process Management*, ser. LNCS, A. Barros, A. Gal, and E. Kindler, Eds. Springer Berlin Heidelberg, 2012, vol. 7481, pp. 335–351.

Energy-Efficient Thread Migration via Dynamic Characterization of Resource Utilization

Claudia Alvarado

Intel Corporation Portland, OR
ca1015@txstate.edu

Dan Tamir and Apan Qasem

Department of Computer Science
Texas State University, San Marcos, TX
{dt19, apan}@txstate.edu

***Abstract*—The affinity of threads with cores in current chip-multiprocessor systems has a substantial impact on the execution time, latency, and power consumption of multi-threaded workloads. Finding an optimal mapping configuration of threads is a significant challenge as it requires detailed knowledge of each thread's demands for shared system resources. This paper describes a software-based strategy that makes judicious thread migration decisions founded on careful inspection of dynamic resource utilization. The main novelty of the system reported in this paper is the extensive utilization of hardware performance counters to develop a set of synthesized metrics that capture resource contention among co-running threads. Experimental results with a set of contemporary parallel workloads, show that the system can achieve significant improvements in power consumption and performance over the default scheduling heuristics implemented in the Linux kernel.**

Keywords–energy efficiency; thread scheduling; workload characterization.

I. INTRODUCTION

The proliferation of portable wireless devices as well as the rapid growth of high-performance server farms and data centers have made power consumption a central point of concern for the entire computing industry. Excessive and unbalanced power consumption of computing systems has a direct economic and environmental impact in the form of high energy bills and large carbon footprints. Additionally, power consumption impacts the computing industry in several indirect ways by adverse effect on device reliability, requiring expensive packaging, and causing irreversible damage to semi-conductor devices. As the industry moves towards the exascale era and the magnitude and volume of computing devices continue to grow, it is clear that power consumption is a dominant metric in the design of computing systems. Several strategies for power-reduction have emerged in response to this challenge in recent years. Several, researchers have focused on hardware techniques such as developing new energy conserving components while others have emphasized software strategies that aim to exploit existing hardware in an energy-efficient manner. This paper focuses on software strategies addressing the problem of determining suitable thread placement policies that improve the energy efficiency of multi-threaded workloads without a commensurate sacrifice in performance.

Power consumption on current chip-multiprocessor (CMP) architectures is influenced by numerous factors including number of cores/threads in the implementation, core frequency application characteristics (e.g., arithmetic intensity), data locality, and cache topology. Because CMP architectures share resources among processing cores, the placement of threads or *thread affinity* can significantly impact the execution of a multi-threaded workload. On one hand if two threads contend for a particular shared resource (e.g., two floating point (FP) intensive threads running on the same hyper-threaded core) it can lead to performance degradation and increase power consumption. On the other hand, a favorable utilization of a shared resource (e.g., shared cache through inter-thread data locality) can result in power-performance benefits. Consider the results of running a parallel workload on an Intel Quad-core system as presented in Figure 1. The workload consists of four parallel applications, executed with five different affinity configurations. The numbers reported are normalized with respect to the default OS-enforced affinity. Significant variations in power-performance are observed for the different configurations. Although the best choice for power and performance coincides with configuration *aff1*, the choices diverge for subsequent points. This makes it imperative that the scheduler considers power-performance trade-offs when making thread placement decisions.

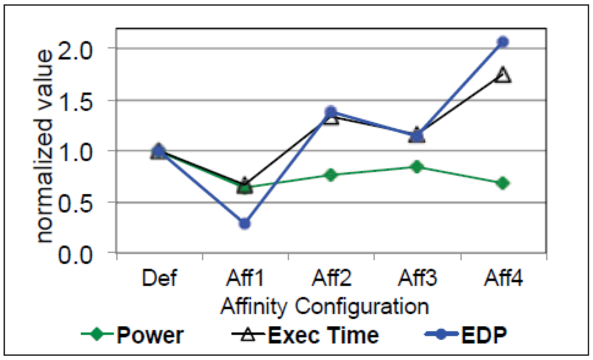


Figure 1. Impact of thread affinity on power, execution time and EDP

Although thread affinity is an important factor in power-performance-based optimizations, developing a scheduling heuristic that works well for different workloads is a challenging task. The difficulty arises from three main sources:

1) *Characterization of resource usage:* The demand and utilization of resources varies across workloads and across programs in a given workload. An intelligent scheduler needs a

ISBN: 978-1-61208-432-9

mechanism that allows monitoring resource usage across the system and characterizing the utilization as either favorable or harmful. Characterization can involve determining inter-thread and intra-thread data locality and arithmetic intensity. This proves particularly problematic for an OS-level scheduler as it needs to extract information without the advantage of source code analysis. Furthermore, the scheduler has to relate resource usage to overall power consumption, a task that is particularly difficult because of the lack of availability of dynamic power consumption data.

2) *Dynamic and fine-grain system monitoring:* The scheduler must collect dynamic measurements at a fine granularity; generally at an interval that coincides with an OS-enforced scheduling quanta (referred to as *slice*). Moreover, this task has to be performed in a non-intrusive manner in order to reduce the impact on the execution of the workload.

3) *Inferring patterns in execution:* In many situations, a small change in the system does not necessitate a change in the placement policy. For example, transient processes often occupy cores for short periods without a major impact on the overall execution time or energy efficiency of the workload. The scheduler needs to be aware of such artifacts and make placement decisions only when a broad pattern change has been detected. For enforcing such a policy there is a need for dynamic feedback as well as for a method of maintaining historical data from which, execution patterns can be inferred.

In this work, we describe an affinity-driven meta-scheduler that addresses each of the above issues. The scheduler operates in the user-space as a runtime system and works in concert with the operating system. Central to our approach is the systematic and extensive utilization of hardware Performance Monitoring Units (PMUs). Today’s commodity processors feature a large collection of PMU counters and registers with advanced capabilities that can provide a wealth of information about system performance. Although the use of PMUs is commonplace in application performance tuning in the high performance (HPC) domain, their use in operating system tasks such as mapping and scheduling is very limited. We leverage these performance counters and construct a set of models that synthesizes PMU counters values to provide insight into performance and energy related issues arising from the execution of a multi-threaded workload. Specifically, we use our PMU-based synthesized metrics to detect performance bottlenecks and power anomalies caused by contention of a shared resource or the under-utilization of a private (exclusive to a core) resource or computational units. We include, in the system, a mechanism to probe the PMU counters, derive the synthesized metrics at fixed intervals, and maintain historical data. The collected data is used as feedback for a novel greedy heuristic-based scheduling algorithm that makes dynamic affinity decisions to improve energy-efficiency and performance. Additionally, the framework facilitates the generation of training data and the integration of machine learning algorithms in scheduling decisions.

This paper makes the following contributions:

- It provides a low-overhead dynamic mechanism for detecting resource contention, sharing, and under-utilization on current multicore architectures.
- It provides new insight about power consumption and utilization of resources and demonstrates how these relationships can be exploited using a novel affinity-based power-aware scheduling heuristic.
- It presents the implementation of a portable meta-scheduler for Linux whose installation requires no modification to kernel code and no instrumentation of application binaries.

The remainder of this paper is organized as follows: Section II discusses related power-aware thread scheduling work. Section III provides an overview of the meta-scheduler framework. Section IV presents the synthesized metrics for resource characterization. Section V presents experimental results and Section VI includes conclusions and proposals for future research.

II. Related Work

Much of the work in scheduling for power has focused on developing runtime strategies that aim to find an optimal schedule for a single parallel application [1][5][8][11]. General strategies for power-aware scheduling are less common [4][7].

Bautista *et al.* present a power-aware scheduler that aims to minimize power consumption while respecting task deadlines in real-time applications [2]. Wierman *et al.* provide theoretical bounds on dynamic voltage and frequency (DVFS) based scheduling techniques [10]. They show that in terms of performance and power, a static DVFS scheduling strategy works as well as a dynamic strategy. However, a dynamic strategy can yield benefits when the objective is to improve system reliability. Kashif *et al.* propose a Priority-based Multi-level Feedback Queue Scheduler (PMLFQS) for mobile devices. PMLFQS is a work-conserving algorithm that uses different central processing unit (CPU) speeds to minimize the overall energy consumed by the CPU for each task [6].

Zong *et al.* have proposed two scheduling algorithms for scheduling parallel applications on large clusters. Their framework utilizes a precedence-constrained task graph of the application to be scheduled and emits a schedule that is predicted to be most energy-efficient [11]. Teodorescu et al. present a power management algorithm that takes into account variations in voltage and frequency among cores and attempts to improve performance within a given power envelope [9].

Merkel et al. develop heuristics that schedule threads according to a resource sharing utility. They combine these algorithms with DVFS techniques and evaluate their strategy on a workload with homogeneous sharing patterns. They demonstrate significant reduction in the Energy Delay Product (EDP) [7]. Boyd-Wickizer et al. propose a technique that operates at the level of objects and migrate threads from core-to-core depending on the data structures they access [4]. Bringing threads closer to the data reduces memory latency [4].

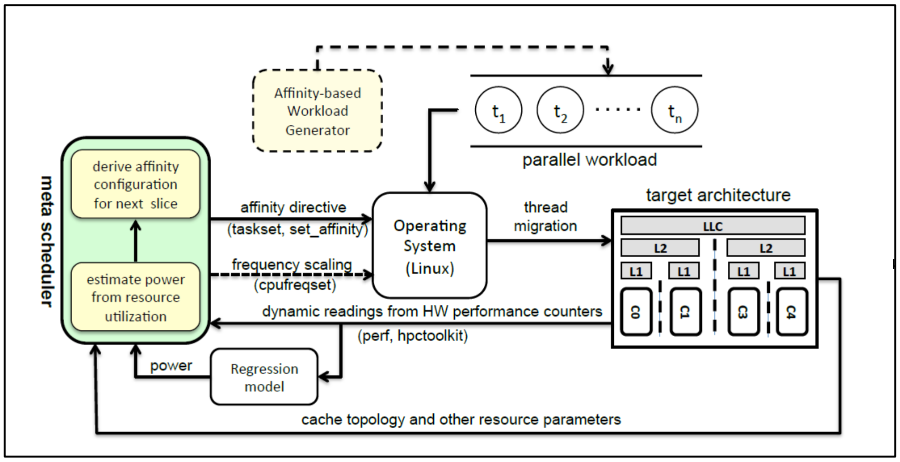


Figure 2. Meta-Scheduler Framework Overview

III. The Meta-scheduler Framework

Figure 2. Provides an overview of the framework. The central component is the meta-scheduler, which executes as a run time system. We have developed an API for communication between the meta-scheduler, the operating system, and the underlying hardware. The API allows the scheduler to probe and measure PMU counters, perform thread migration, and set core frequency. Special attention is given to ensure that each of these tasks are accomplished with low overhead. Next, we briefly discuss the implementation of each of these interfaces.

Hardware performance counters can provide detailed system information during workload execution. Software for probing this hardware has matured and there are several related tools such as *Intel Vtune*, *PAPI*, *HPCToolkit*, and *Oprofile*. Most of these tools, however, are primarily designed for single application tuning and those that do system-level profiling (e.g., *Oprofile*) carry significant overhead. Because our scheduler has to run at a low overhead we opted to implement our own interface that directly reads the PMU register values from device files.

The Linux *taskset* utility is used to set or change the affinity of a particular thread. The utility allows specifying the subset of cores that an application can be run on. It enforces a hard affinity on thread execution, which means that the OS always honors the affinity set by *taskset*.

It is difficult to determine dynamically in an inexpensive way if there is performance change in power demands. A run-time system that has access to PMU counters can monitor certain *trigger effects* that provide intuition into performance and power issues of running programs. This insight when modeled into a scheduling algorithm can yield significant gains.

In this section, we provide descriptions of three such resource-utilization metrics that are employed in our system. In the discussion that follows, we use the hardware counter names from the Intel Core micro-architecture. Similar counters are available on the AMD Barcelona and the IBM Power 7.

A. Core Utilization

Utilization of cores by different threads in a parallel system, is a key indicator of performance. Generally, a system running with a balanced load, where all cores are performing a similar amount of work, yields maximum concurrency and consequently leads to better performance. The utilization of cores has special significance when considering power consumption. With core gating it may be beneficial to consolidate the load into a subset of the cores while power gating the remaining cores. Regardless of the end objective, it is important to consider the workload balance while making thread mapping and scheduling decisions. These decisions, however, require accurate and dynamic measurements of core utilization at every time slice.

Current methods of measuring CPU utilization (e.g., Linux *top* utility) are not suitable for multicore architectures with multi-level caches and non-uniform memory accesses. In our framework, we use a set of counters to accurately estimate the amount of work done by each core. At each time slice we inspect the counters that provide the number of cycles a core is busy and the elapsed time (i.e., slice length). The number of elapsed cycles is a function of the clock frequency and elapsed time. However, because the core frequency can be modified during a given slice (e.g., via Intel's *Turboboost*), we use the *cpugovernor* to determine the current operating frequency of each core. The obtained frequency is used to determine the total number of elapsed cycles and the ratio of the busy cycles to elapsed cycles provides the core utilization.

B. Cache Behavior

The quality of cache utilization depends on intra and inter-core data locality. Yet, it is difficult to determine program locality without *a-priori* information. Nevertheless, by the examination of a set of PMU counters it is possible to determine favorable and non-favorable cache behavior.

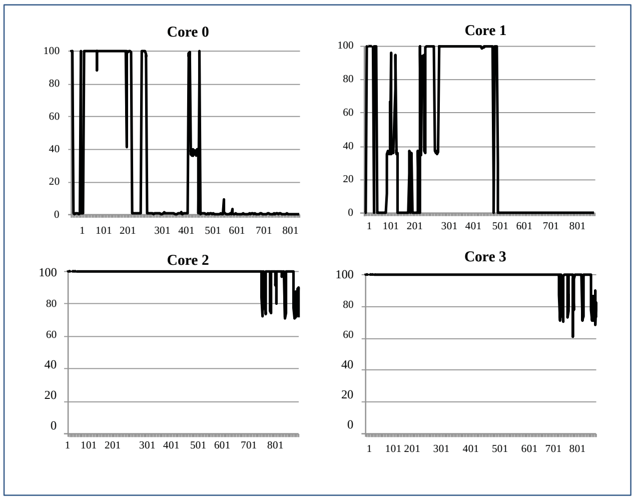


Figure 3. Per-core utilization for four-program workload broken down by core

If two threads share a cache and they have no shared data locality, then cache utilization is generally determined by their respective *working-sets* i.e., the amount of data accessed by each thread repeatedly. If the working set of the two programs exceeds the capacity of the shared cache then threads incur numerous misses in short succession. Applications do not necessarily access working sets during the entire execution. Therefore, the condition that needs to be checked is if both threads hit their respective working sets during the same time interval and exceed the capacity of the shared cache. This can be achieved by tracking per-core cache miss rates for the shared caches. There is contention in a shared cache if the average miss rate for shared cache in the last *k* intervals is significantly greater than the average miss rate of the same cache for the previous *j* intervals. A significant increase is determined by using a tolerance value as a tunable parameter. Since any applications can have multiple working sets that correspond to different caches, it is important to inspect contention at multiple levels.

Inter-core and inter-thread locality can have an impact on performance. If two threads have shared access to data and they are mapped to a set of cores that share cache then both execution time and power consumption benefits due to reduced cache misses. On the other hand, if two threads have no locality and they compete for cache space, then increase in cache misses reduces performance. Our system utilizes a set of performance counters to determine both favorable and unfavorable sharing of cache.

C. Computation Unit Utilization

System units can be shared at the hardware and thread level. For example, with hyper-threading two software threads running on the same core can share the FP unit. In this case it would be prudent to place threads that are FP-intensive onto different cores. Our system accounts for only one such resource, namely the on-chip FP units. However, as the system-on-chip (combined *GPU-CPU*) architectures become more prevalent, tracking utilization of other shared computational units will become more important.

V. Evaluation

In this section, we present experimental results that illustrates the significance of resource characterization metrics in the energy-efficient execution of parallel workloads. We then show their effectiveness in making thread migration decisions with a greedy algorithm.

A. Experimental Setup

1) Platforms: the platform used is an Intel quad-core system that contains two Core 2 Duo processors sharing an L2 cache. Each core has a private L1 cache. The system runs Linux kernel 3.0. Under our adaptive thread migration policy the *cpugovernor* is set to custom. This eliminates interfere with the heuristics. When using the Linux strategy the *cpugoverner* is set to default.

2) Benchmarks: We evaluate our strategy on a variety of workloads generated from the *PARSEC* benchmark suite [3]. The suite includes a collection of multi-threaded programs with varying demands for system resources and contains data-, task-, and pipelined parallel applications. Each workload is formed from a subset of the *PARSEC* applications.

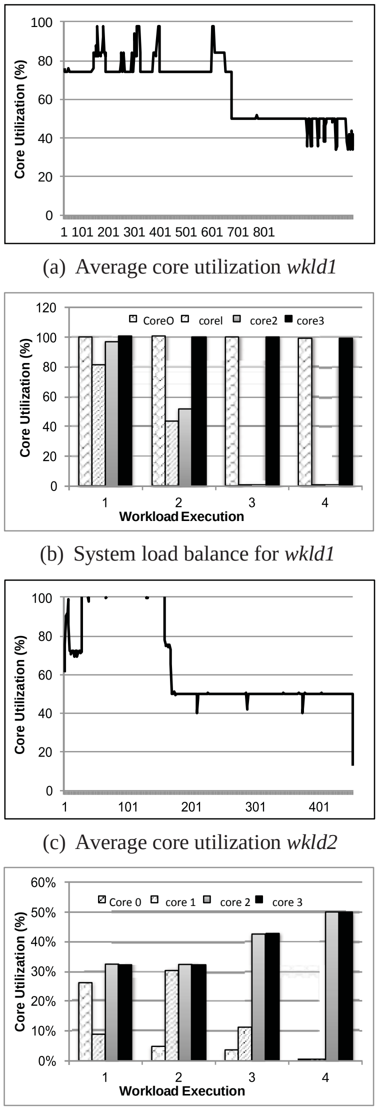
B. Load Balancing

First, we examine the way that workload characteristics and their affinity configurations impact the core utilization metric and the overall load balance of the system. Figure 3 shows the average core utilization of individual cores for *wkld1*, consisting of *canneal*, *streamcluster*, *blackscholes*, and *freqmine* [3]. The default migration policy of Linux is used in executing the workload. Significant variations in utilization of each core are observed throughout the execution of *wkld1*. In particular, cores 0 and 1 exhibit poor utilization during the first 400 time slices, while remaining under-utilized towards the end of the execution. Figures 4(a) and 4(b) show the average core utilization and the load balance of the system during four segments of execution. The system is close to a balanced state for only a small fraction of the time. More significant, however, is the fact that the average core utilization metric provides a clear indication concerning the balance of the system. The low utilization during times slices 300-400 would be a trigger for a smart scheduler and adjust the affinity to achieve better balance. Figures 4(c) and 4(d) present average core utilization and load balance information for a second workload (*wkld2*) that consists of *fluidanimate*, *canneal*, *streamcluster*, and *dedup* [3]. Interestingly, although *wkld2* contains two of the same applications as *wkld1*, we observe significant differences in the average core utilization. Even for this workload the average core utilization is a good indicator for the system load balance.

C. Cache Behavior

Figures 5 and 6 show last level cache (LLC) misses for *wkld3*, which consists of *raytrace*, *swaptions*, *streamcluster*, and *dedup* [3]. We observe that for both affinity configurations there is considerable fluctuations in the cache miss rates. These fluctuations, however, do not follow the same pattern for the two different configurations.

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-432-9



(a) Average core utilization *wkld1*

(b) System load balance for *wkld1*

(c) Average core utilization *wkld2*

(d) System load balance for *wkld2*

Figure 4. Variations in average core utilization and load-balance

This demonstrates the way that the cache miss rate can be impacted by the choice of affinity. The most interesting aspect of these results are the spikes in cache miss rates observed at various intervals (e.g., core 3 for *wkld2* at interval six). These sudden spikes can have a negative impact on performance and power consumption. To ameliorate the ill-effects of these spikes, the scheduler must have information at time slices boundaries, as provided by our framework.

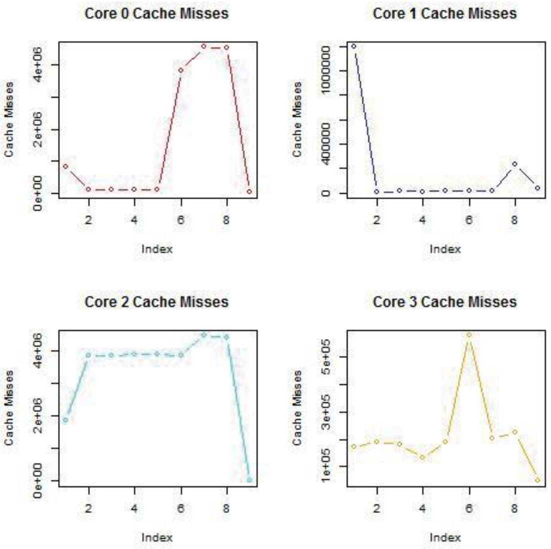


Figure 5. Core-level breakdown of LLC misses for *wkld2* with default affinity

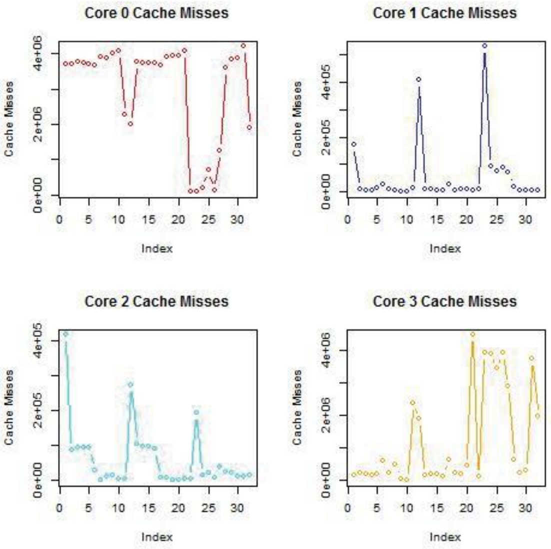


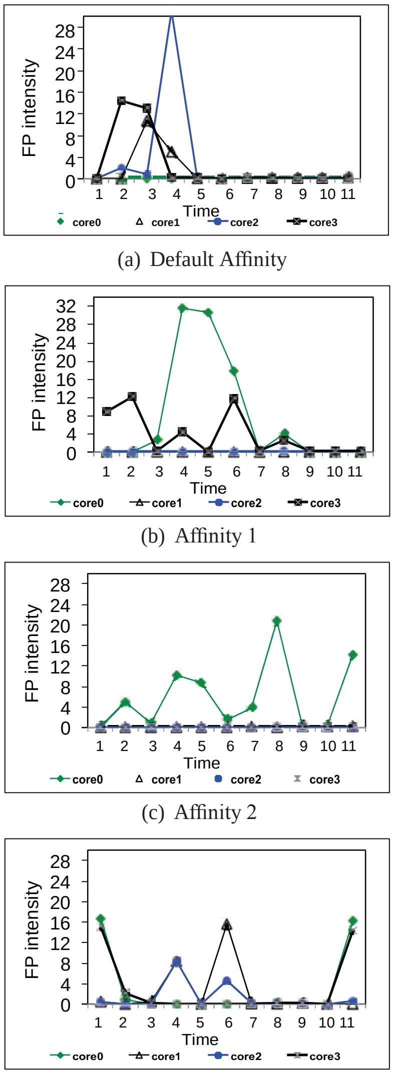
Figure 6. Core-level breakdown of LLC misses for *wkld2* with affinity configuration, aff2

D. Computational Units

Figure 7 shows variations in arithmetic intensity for different affinity configurations for *wkld4*, which consists of *canneal*, *streamcluster*, *facesim* and *x264* [3]. Considerable variation in arithmetic intensity exists when different affinity policies are used. The *aff3* strategy shows the most balance in distribution of FP operations across the cores. Both *aff1* and *aff2* produce several spikes in FP-activity on core 0. For the default affinity, most FP operations are packed towards the beginning of the workload execution; but they are under-utilized later on.

E. Evaluation with a Greedy Algorithm

We have implemented an adaptive thread migration algorithm that exploits resource characterization metrics and makes decisions based on a greedy heuristic. At each time slice the algorithm inspects the system, collects measurements, and tracks all currently running processes. A history table is used to store the resource usage data from the last *k* intervals. At each time slice the algorithm makes a decision on whether to change the current affinity in order to improve the load balance, cache sharing, and FP unit utilization. All decisions are weighed against the predicted power consumption for the next time slice.



(a) Default Affinity

(b) Affinity 1

(c) Affinity 2

(d) Affinity 3

Figure 7. Variations in arithmetic intensity for different affinity configurations *wkld4*

TABLE I. ENERGY EFFCIENCY WITH GREEDY HEURISTICS

Workload	Power (W)		Exec. Time (s)		Energy (K Joules)	
	Linux	Greedy	Linux	Greedy	Linux	Greedy
wkld1	38.37	27.98	355	797	13.62	22.30
wkld2	38.25	32.93	306	277	11.71	9.12
wkld3	39.34	25.26	307	204	12.07	5.15
wkld4	28.16	26.59	10	13	0.28	0.34

Table I presents results of applying our algorithm on the four different workloads discussed earlier. We observe that in almost all the cases, the greedy algorithm outperforms the Linux scheduler in terms of energy dissipation and execution time. Particularly compelling is the situation with *wkld3*, where our greedy heuristic results in a 35% reduction in power consumption. Overall, on average, the greedy heuristic yields a 2% reduction in energy, 22% reduction in power consumption, and 32% increase in execution time.

VI. CONCLUSIONS

This work presents an energy-efficient thread migration strategy that is based on characterization of resource usage. We have identified a set of synthesized metrics that provide key insight into the execution behavior of parallel workloads running on contemporary multicore architectures. The experimental results show that core utilization, cache contention, and use of FP units can impact the execution and power consumption in intricate ways. We develop a greedy algorithm that exploits these synthesized metrics to significantly outperform the Linux scheduler both in terms of performance and energy efficiency.

In the future we plan to further the research and evaluate several rescheduling mechanisms. Additional avenue is to include additional shared resources.

VII. ACKNOWLEDGMENTS

Financial support for this work was provided by the Semiconductor Research Consortium (SRC) under contract no. 2011-HJ-2156 and the National Science Foundation through awards nos. CNS-1305302 and CNS-1253292.

REFERENCES

[1] I. Ahmad, R. Arora, D. White, V. Metsis, and R. Ingram, “Energy-constrained scheduling of dags on multi-core processors,” in S. Ranka, S. Aluru, R. Buyya, Y. C. Chung, S. Dua, A. Grama, S. K. S. Gupta, R. Kumar, and V. V. Phoha, editors, Contemporary Computing, volume 40 of Communications in Computer and Information Science, pp. 592–603. Springer Berlin Heidelberg, 2009.

[2] D. Bautista, J. Sahuquillo, H. Hassan, S. Petit, and J. Duato “A simple power-aware scheduling for multicore systems when running real-time applications,” in the IEEE International Symposium on Parallel and Distributed Processing, pp. 1-7 2008.

[3] C. Bienia, S. Kumar, J. P. Singh, and K. Li, “The PARSEC Benchmark Suite: Characterization and Architectural Implications,” in the Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques, pp. 72-81, 2008.

[4] S. Boyd-Wickizer, R. Morris, and M. F. Kaashoek, “Reinventing scheduling for multicore systems,” in the Proceedings of the 12th conference on Hot topics in operating systems, pp. 21-22 2009.

[5] Y. Guo, J. Zhao, V. Cave, and V. Sarkar, “Slaw: A scalable locality-aware adaptive work-stealing scheduler,” in the IEEE International Symposium on Parallel Distributed Processing, pp. 341-342 2010.

[6] M. Kashif, T. Helmy, and E. El-Sebakhy. “A priority-based mlfq scheduler for CPU power saving,” in the Proceedings of the IEEE International Conference on Computer Systems, pp. 130-134, 2006.

[7] A. Merkel, J. Stoess, and F. Bellosa, “Resource-conscious scheduling for energy efficiency on multicore processors,” in the Proceedings of the 5th European conference on Computer systems, pp. 153-166, 2010.

[8] Q. Tang, S. K. S. Gupta, and G. Varsamopoulos, “Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach,” IEEE Transactions on Parallel and distributed systems, vol. 19, pp. 1458–1472, 2008.

[9] R. Teodorescu and J. Torrellas, “Variation-aware application scheduling and power management for chip multiprocessors,” in the Proceedings of the 35th Symposium on Computer Architecture, pp. 353-374, 2008.

[10] A. Wierman, L. Andrew, and A. Tang. Stochastic analysis of power-aware scheduling,” in the proceedings of the 46th Conference on Communication, Control, and Computing, 2pp. 23-26, 2008.

[11] Z. Zong, A. Manzanares, X. Ruan, and X. Qin, “Two energy-aware duplication scheduling algorithms for parallel tasks on homogeneous clusters,” IEEE Transactions on Computers, pp. 360-374, 2009.

Evaluating Neural Network Methods for PMC-based CPU Power Prediction

Mario Gutierrez, Dan Tamir, and Apan Qasem
Department of Computer Science
Texas State University
San Marcos, TX
{mag262, dt19, apan}@txstate.edu

***Abstract*—Emphasis on energy efficient computing has established power consumption, as well as energy and heat dissipation as determinant metrics for analyzing High performance computing applications. Consequently, optimizations that target High performance computing systems and data centers have to dynamically monitor system power consumption in order to be effective. Current architectures are exposing on-chip power sensors to designers and users. The general state of power measurement tools across different architectures, however, remains deficient. Recent research has shown that first-order, linear models can be effectively used to estimate real-time power consumption. This paper describes a neural-network based model for fine-grain, accurate and low-cost power estimation. The proposed model takes advantage of the wide array of performance monitoring counters available on current systems. We analyze the prediction capability of the model under various scenarios and provide guidelines for feature selection for other machine learning models for estimating power consumption on future architectures.**

Keywords–energy efficiency; power consumption; workload characterization, performance counters.

I. INTRODUCTION

The proliferation of portable wireless devices, along with the rapid growth of high-performance server farms and data centers, have made energy efficiency a central concern for the entire computing industry. Poorly managed and unbalanced power consumption of computing systems has direct economic and environmental impact in the form of high energy bills and large carbon footprints. Furthermore, it leads to device reliability degradation and high chip packaging costs. In recent years, numerous strategies for reduction in power usage have emerged. Current research is focused on hardware techniques such as developing new energy conserving components and on software strategies that aim to exploit existing hardware in an energy-efficient manner.

Most software techniques rely on methods for estimating system power in making optimization decisions. Several architectures provide a mechanism for measuring power directly. But for architectures that do not expose these metrics, obtaining system power involves attaching an external device or running costly simulations [1]. Neither method is suitable for software-based techniques that need to react to changes in power usage and make real-time decisions to conserve energy. Furthermore, even on platforms where the power counters are available, measurements incur a high overhead.

An efficient solution can be constructed via the reuse of hardware devices. One of the most important elements used in previous research is the set of built-in Performance Monitoring Counters (PMCs). These are registers built into the Central Processing Unit (CPU) and/or into the Performance Monitoring Unit (PMU) that can track performance-related information such as instruction counts, cache misses, and resource stalls.

Several models that correlate program behavior, captured via PMCs, to system power consumption have been proposed. In the body of research on modeling of CPU power consumption using PMCs, the dominant trend is to use linear regression. A linear framework is useful for understanding the relations between and importance of the independent variables used, but it is very rigid and requires tailoring to the specific problem.

This article presents an analysis of the power prediction behavior of a linear regression model from various perspectives and compares it to a neural network model. The purpose of this is to discover the underlying patterns of these models on the specific task of estimating and predicting power through the PMU hardware. This knowledge can serve as a guide for researchers for further investigation of power models.

The paper is organized as follows: Section II discusses related work. Section III details the research methodology, the experimental setup, the experiments, and the results obtained. Section IV includes result evaluation and Section IV presents conclusions and proposals for future research.

II. RELATED WORK

The use of PMCs for energy-aware applications has been heavily researched over the past few years. One of the first links between PMCs and power consumption has been pointed to by Bellosa [2]. In his paper, Bellosa has demonstrated the high correlation of PMCs to power consumption, and has presented strategies for energy-aware scheduling.

Contreras et el. have used PMCs in a linear model for the prediction of CPU (and memory) power consumption [3]. Singh et al. demonstrated the use of PMCs for power-aware thread scheduling [4]. Nagasaka et al. have been able to achieve good results by using PMCs to estimate GPU energy consumption [5]. Stockman et al. has examined machine learning techniques for the prediction of memory power usage [6]. Rodrigues et al. have shown that a power estimation model trained on one CPU can be used with reasonable accuracy on other CPUs with similar micro- architectures [7]. Lee et al. have investigated the use of PMCs for estimating micro-architectural components temperature [8]. Cavazos et al. have presented work on using PMCs to determine the best compiler optimization settings [9].

While these papers demonstrate the importance of PMCs for CPU performance evaluation, the use of PMCs for power estimation and prediction is concentrated on regression models. Our literature survey did not identify reference to the usage of neural networks for this process.

ISBN: 978-1-61208-432-9

III. METHODOLOGY

This section includes details concerning the setup, data collection, and models used in the reported experiments. This information is useful as a point of reference concerning the machines and models in the experimental section as well as for implementing similar experiments on additional machines.

A. *Selecting PMC Events*

In the initial stages of the project, twelve PMCs of an Intel-based system have been considered for use in the models. For the final models, the five with the highest correlation value and the resource stalls counter have been chosen. Similar PMCs to those six have been located for the AMD machine and for the Power PC architecture.

B. *Generating Training Data*

To generate the training data for the learning models we have utilized a workload generation script that creates a large number of parallel workloads with diverse characteristics. The script selects a subset of the *PARSEC* programs and executes them by varying execution parameters such as the number of threads and data set size [1][3][7]. During each run of a workload, the PMC events listed in Table I are probed at a fixed interval. The power consumption values are recorded using the available power sensors on the *Sandybridge* architecture [1][3][7]. We have used an interval of 10 seconds.

The *Watts-Up-Pro* power meter is used for collecting power samples from the AMD machine since that machine does not have power sensors [7]. Initially, unnecessary background programs have been disabled, the workload script has been initiated, and peripherals have been disconnected. The workload has generated PMC event samples with an added time stamp component to assist in the synchronization with the power samples. Power samples have been collected every 2 seconds, and the PMC samples have been collected every 8 seconds (16 total as a sample is collected in two steps).

After the workload script has completed, the power samples have been collected from the memory of the PMU. To create the final dataset, the four power samples leading up to the elapsed time per PMC half-sample are found. Then, the eight power samples per PMC sample are averaged and used in the final dataset. This method has proved as highly reliable. The workload power patterns have been found to be well defined and prediction on the set has been very accurate.

C. *Linear Regression Model*

In these experiments, a straightforward multivariate linear regression model is used. The model is expresses via the following equation:

$$Ypwr = a1 \times p1 + a2 \times p2 + \ldots + a6 \times p6 \quad (1)$$

Where pi are the PMC counter values, ai are the coefficients to be trained/identified, and $Ypwr$ is the power value. The model starts with $Ypwr = a1 \times p1$ and a term, $ai \times pi$ is added at each step. The *lm* training function from R is used.

D. *Neural Network Model*

In contrast to the linear model, a neural network is more flexible, but more difficult to train. Additionally, because the weights are randomly initialized and the training only finds local minima, there is a problem with the results variance. There are several types of neural networks; for the reported experiments, a multi-layer, feed-forward structure utilizing the *neuralnet* implementation from R has been used. The default resilient back-propagation with weight backtracking procedures are used for training the model [10]

A neural network model consists of an input layer, one or more hidden layers, and an output layer. Our neural network model has one node in the input layer for each PMC event, and one node for the power value in the output layer. There are two hidden layers, one layer with 5 nodes and a second layer with 3 nodes. Originally, the model had only one layer, however the performance has been deficient. The addition of an extra layer has achieved satisfactory accuracy. The final values, which have yielded the local minimum, have provided significant improvement over the linear regression model.

Every node in a layer of the neural network model is mapped to every other node in the next layer. A mapping from one node to another can be represented by a weighted edge and finding these weights is the goal of training the network. The values of each node are calculated using: the values from the previous layer, the weights, and a squashing function. As an example, let v_{jm} be the j^{th} node in layer m, and let v_{in} be the i^{th} node in the previous layer n, let $w_{i,j}$be the weight from node i to node j, and let $sqsh$ be the sigmoid function. Then,

$$v_{jm} = sqsh\left(\sum_{i=1}^{n}(v_{in} \times w_{i,j})\right) \quad (2)$$

The hidden and output layers have an additional bias node whose value is always 1.

IV. EVALUATION

This section presents and discusses the results of various experiments. Each experiment examines a particular aspect of power prediction using the data, models, and setup described in the previous section.

A. *Platforms*

Table II provides details on the three systems that served as the evaluation platforms. In the rest of the paper, we refer to these platforms using the names listed in the header row. We have included both AMD and Intel-based systems in our experiments as the PMC units differ significantly.

B. *Effect of Sample Size*

The first experiment has explored the effect that the number of power samples has on the amount of prediction error. The purpose of this experiment is to determine the benefit of using more power samples for training. Figure 1 provides the results of this experiment.

At the beginning of the procedure, the samples of the dataset are normalized and shuffled. We have chosen to run the experiment for 24 trials of increasing sample size.

TABLE I. PERFORMANCE COUNTER EVENT NAMES PER MACHINE

Phenom	*Sandybridge*
CPU_CLOCKS_UNHALTED	UNHALTED_CLK_CYCLES
INSTRUCTIONS_RETIRED	INSTR_RETIRED_ANY
UOPS_RETIRED	UOPS_RETIRED_ALL
DISPATCHED_FAST_FPU	FP_COMP_OPS_EXE_X87
BRANCH_MISPREDICT_RETIRED	BR_MISP_RETIRED_ALL_BRANCHES
DISPATCH_STALLS	RESOURCE_STALLS_ANY

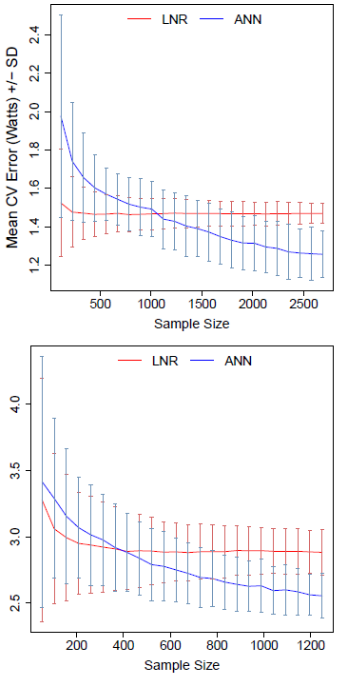


Figure 1. Sample Size Result on *Phenom* (l) and *Sandybridge* (r)

The next step is used to determine the number of samples Needed for increasing the dataset between each trial. After this, the dataset of_ each trial is split into 6 disjoint partitions for cross-validation (6-fold). Multiple runs of the procedure produce graphs with variability that makes it difficult to discern a clear pattern. To emerge the underlying pattern, the procedure has been repeated 50 times and the values are averaged. Both of the graphs for this experiment used the same procedure. One result used samples of the *Phenom* dataset [7], and the other used samples of the *Sandybridge* dataset.

The experiment results are interesting in several ways. First, the results are very similar despite the fact that each run used a dataset from different architectures. This is representative of the similar nature of the datasets, despite the fact they came from different architectures. Second, the linear regression model reaches its best performance for a small number of samples.

Third, the neural network eventually outperforms the linear regression model, but at different numbers of sample size. We speculate that the neural network overtakes the linear model at different points both because of the amount of noise in the *Sandybridge* data, and the inherent differences that are due to differing architectures.

The noisiness of the *Sandybridge* data has the larger effect as it prevents the linear regression model from converging to a lower error. Thus, the intersection point moves leftward. This assumes that the neural network can better handle noisy data.

In general, not much improvement in accuracy is seen between the models by the maximum amount of samples used. It is necessary, however, to have at least one thousand samples for obtaining reasonable variance. It is also expected that the neural network will continue to achieve slightly better accuracy for larger dataset sizes. A good strategy for quick power modeling is to collect one thousand samples and use a linear regression model. If accuracy the observed accuracy is not sufficient, a neural network model with a few thousand samples is recommended. We expect future collected datasets of the same type as used in this paper to behave similarly.

C. *Effect of the Number of PMCs*

The next experiment has examined the way that the performance of a power model behaves as more PMCs counters are included. Figure 2, depicts the experiment results. At the start of the procedure, the samples of the dataset are normalized and shuffled. The dataset at each trial is then split into 6 disjoint partitions for cross-validation. This procedure is conducted on each dataset to obtain the respective results.

TABLE II. EVALUATION PLATFORMS

	Core	*Sandybridge*	*Phenom*
Cores	4	6 (12 logical with HT)	4
Processor	2.40 GHz Intel Core 2 Quad	2.0 GHz Intel Xeon	1.00 GHz AMD Phenom
L1	32 KB (private)	32 KB (private per physical core)	64 KB (private)
L2	2 × 4 MB (shared 2 cores)	256 (private per physical core)	512 KB (private)
L3	none	15 MB (shared all)	2 MB (shared)
Compiler	GCC 4.8.2 -O2	GCC 4.8.2 -O2	GCC 4.8.2 -O2
OS	Ubuntu 14.04.1	Ubuntu 14.04.1	Ubuntu 14.04.1
Kernel	3.13	3.13	3.13

TABLE III. PMC EVENT NAMES IN ORDER OF ADDITION FOR EXPERIMENT 4.2

	Phenom	*Sandybridge*
1	INSTRUCTIONS_RETIRED	UNHALTED_CLK_CYCLES
2	UOPS_RETIRED –	INSTR_RETIRED_ANY –
3	BRANCH_MISPREDICT_RETIRED	UOPS_RETIRED_ALL
4	CPU_CLOCKS_UNHALTED	FP_COMP_OPS_EXE_X87
5	DISPATCHED FAST FPU	BR MISP RETIRED ALL BRANCHES
6	DISPATCH STALLS	RESOURCE STALLS ANY

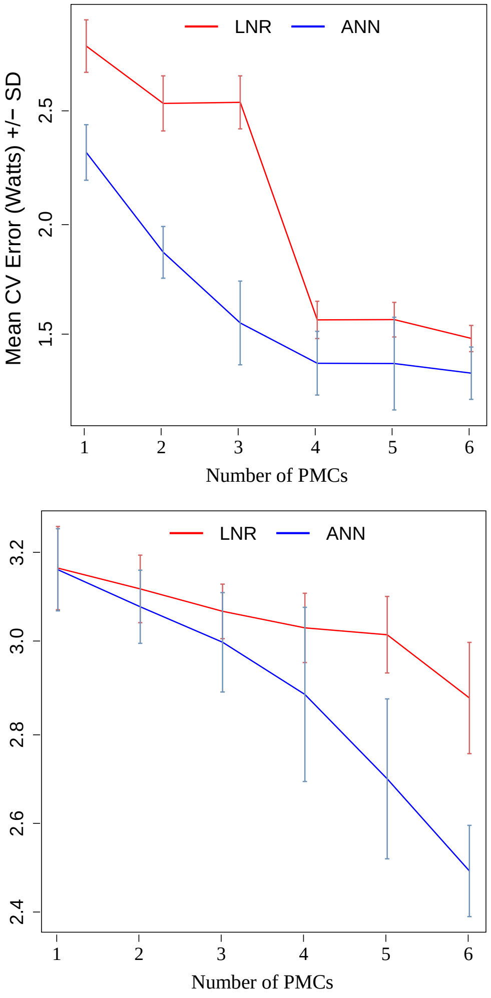


Figure 2. Number of PMCs Result on *Phenom* (l) and *Sandybridge* (r)

The PMCs have been added to the feature vector in order of decreasing Spearman correlation value. This method can be used with any set of PMC events. Table III provides the names of the PMCs in the order of insertion. The results from this experiment did not exhibit similar patterns.

For the *Sandybridge* dataset, both models have performed very similarly with one feature. Each subsequent feature improved performance for each model, showing stronger improvement for the neural network model. For the *Phenom* dataset, the neural network outperformed the linear regression model up until the fourth feature, after which the difference in performance became smaller. A few phenomena have been observed from these results. First, having more features improves accuracy. Second, a neural network model slightly outperforms a linear regression model with the same feature set.

Most machines have built-in PMCs, but only a few performance counters can be sampled at a time. From these experiment, it is easy to see that better results are obtained when more features are sampled. Furthermore, four metrics is a sufficient number of features to sample at a time. PMCs can be collected in alternating sets. This, however, slows down the rate of update in a real-time system by a multiple of the amount of sets. Additionally, there are problems if the PMCs are sampled over long periods of time.

D. Performance on New Workloads

To measure performance on new workloads, rather than unseen samples, the following experiment is performed. With this experiment, the error of the models on unseen workloads is observed. This may give an account the type of workloads that are not well predicted by our set of PMC events. The experiment results are depicted in Figure 3.

In this experiment, both the *Phenom* and *Sandybridge* datasets have been used. Initially, each dataset has consisted of samples collected from 66 different workloads. Both datasets shared a very similar distribution of samples for the respective workloads. The first step in preparing the datasets for the experiment is excluding very small workloads. The next step is trimming down all the remaining datasets (54 data sets) to the size of the sample with the minimum number of samples. This eliminates the likelihood that a large sample will cause a training. Next, for each reaming workload, the models are trained on samples from all other workloads and the prediction error is determined.

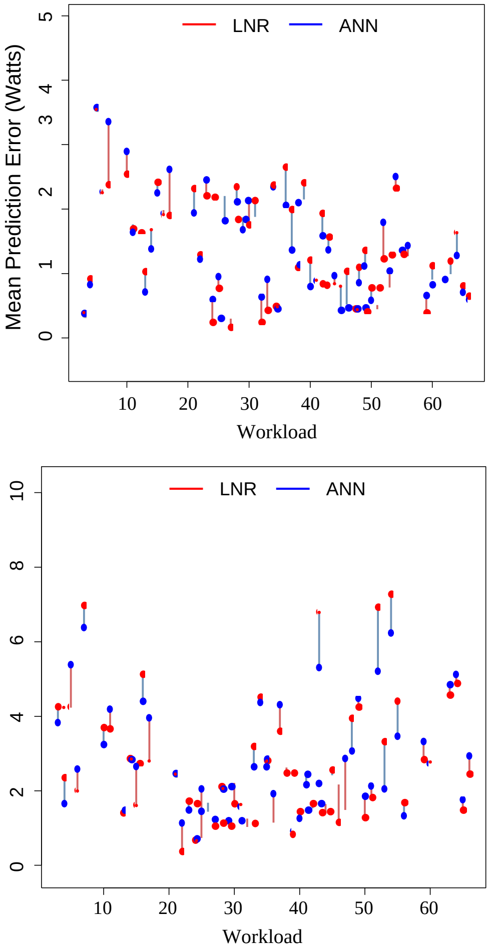


Figure 3. Feature Selection Result on *Phenom* (l) and *Sandybridge* (r)

The procedure is executed 50 times and averaged for the final result. This method is used to produce the resulting graph for the *Phenom* and *Sandybridge* datasets (Figure 3).

The graphs of Figure 3 are structured in a way to display the error per workload as a set of distinct trials, and to accentuate the relative error between the linear regression and neural network. Each point plots the prediction error on a workload after being trained on all the other workloads. There are two points per workload trial, a blue one for the neural network result and a red one for the linear regression result. A line connects a pair and has the color of the best performing model.

The models have very similar performance on *Phenom* 35.2% of the time (less than 0.15 Watt difference). They demonstrate similar performance on *Sandybridge* 27.8% of the time (less than 0.20 Watt difference). With the *Phenom* data, the neural network outperforms the linear regression model 59.3% of the time. With *Sandybridge*, the linear regression model won out 55.6% of the time. On average, *Phenom* had error of 1.42 +/- 0.68 Watts for the linear regression model and 1.40 +/- 0.68 Watts for the neural network model. On *Sandybridge*, the average error is 2.75 +/- 1.60 Watts for the linear regression model and 2.78 +/- 2.37 Watts for the neural network model.

A. *Cross-Architecture Power Prediction*

The final experiment performed discovers whether patterns in power prediction are reasonably similar between different architectures. If this is the case, it would be reasonable to train a predictive model on one architecture and use it for prediction on other machines. The experiment results are included in figure 4.

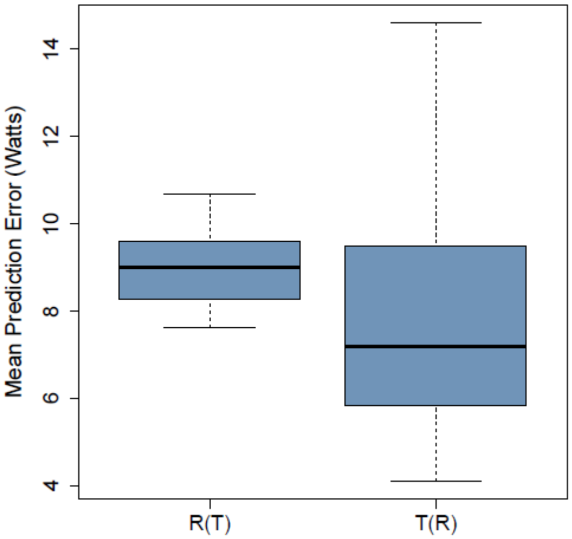


Figure 4. Cross-Architecture Result for Neural Network Model

The procedure for this experiment starts with the normalization of datasets to the range of [0, 1]. After this step, the models are trained on the data from one machine and used for prediction on the other machine. To settle the intrinsic variance in the neural network model, we have averaged the results over 50 experiments.

The linear regression model is able to achieve a prediction error of 8.13 +/- 5.42 Watts when trained on the *Sandybridge* data and used for prediction on *Phenom*. The error is 9.28 +/- 5.32 Watts when trained on *Phenom* and used for prediction on *Sandybridge*. These are moderately acceptable results. The box objects in Figure 4 show the performance of the neural network when trained on *Phenom* and used for prediction on *Sandybridge* ($R(T)$ and $T(R)$). The $R(T)$ performance is about the same as the linear regression model achieved. The $T(R)$ performance had a relatively low median error, but high variance.

In conclusion, the result of this experiment is quite surprising. We did not expect the models to perform as well as they did. However, the performance of prediction between machines with vastly different architectures is still quite poor and not a recommended alternative to the collection of training data from the target machine itself. On the other hand, it is very these results signify that prediction between similar machines could be performed much more accurately. This is in line with the cross-architecture prediction results of [7].

V. CONCLUSION AND FURTHER RESEARCH

Several aspects of power prediction using PMCs are examined in this paper. These are: the effect of the number of power samples used, the effect of the number of performance counters used, the predictive accuracy on unseen workloads, and the prediction accuracy using training data from machines with different architectures.

In the sample size experiment, it is concluded that more samples improved the performance of neural network model consistently. On the other hand, the linear regression model settled onto its best accuracy after only a relatively small number of samples. The immediate conclusion is that using a linear regression model is probably fine for most applications and in cases where only a small number of samples are available. Otherwise, if high accuracy is desired, it would be better to use a neural network with a high number of samples.

The next experiment has examined the way that an increasing number of performance counters affects prediction accuracy. It has been concluded that using more performance counters (starting with the highest correlated counter) generally leads to better accuracy. Four counters has provided sufficient accuracy for both machines tested. The experiment for prediction ability on new workloads showed no obvious patterns between the two machines. This result may indicate that the prediction error of this type of workload, cannot be accurately determined given a set of PMCs. This could be, however, the cause of a difference between implementation of the PMCs on the different machines.

The final experiment has examined the prediction performance of training the models on one machine and using them for prediction on the other machine. The result is not "bad" given that the two machines used are from very different architectures. This shows that there exist common patterns in the collected PMC data between machines. While this would not be useful for very different machines, it could be useful when considering machines with similar architectures. A power-aware scheduling program built on one machine can work well on a similar machine without having to collect any new samples.

In tandem, these experiments highlight useful behavior of power-PMC modeling concentrating on the use of neural networks for this purpose.

In the future we plan to explore additional neural networks and perform additional experiments for PMU counter selection. Additionally, the power estimation and power prediction models will be embedded in a meta-scheduler developed by the research team.

VI. ACKNOWLEDGMENTS

Financial support for this work is provided by the Semiconductor Research Consortium (SRC) under contract no. 2011-HJ-2156 and the National Science Foundation through awards nos. CNS-1305302 and CNS-1253292.

REFERENCES

[1] S. Li, J. H. Ahn, R. Strong, J. Brockman, D. Tullsen, and N. Jouppi, "McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures," in the IEEE/ACM International Symposium on Microarchitecture, pp. 469-480, 2009.

[2] F. Bellosa, "The benefits of event: Driven energy accounting in power-sensitive systems," in Proceedings of the 9th Workshop on ACM SIGOPS European Workshop, pp. 37–42, 2000.

[3] G. Contreras and M. Martonosi, "Power prediction for intel xscale reg; processors using performance monitoring unit events," in the Proceedings of the International Symposium on Low Power Electronics and Design, pp. 221–226, 2005.

[4] K. Singh, M. Bhadauria, and S. A. McKee, "Real time power estimation and thread scheduling via performance counters," SIGARCH Computer. Architecture News, 37(2), pp. 46–55, 2009.

[5] H. Nagasaka, N. Maruyama, A. Nukada, T. Endo, and S. Matsuoka, "Statistical power modeling of GPU kernels using performance counters," in the proceedings of the International Green Computing Conference, pp. 115–122, 2010.

[6] M. Stockman, M. Awad, R. Khanna, C. Le, H. David, E. Gorbatov, and U. Hanebutte, "A novel approach to memory power estimation using machine learning," in the proceedings of the International Conference on Energy Aware Computing, pp. 1–3, 2010.

[7] R. Rodrigues, A. Animalia, I. Koren, and S. Kundu, "A study on the use of performance counters to estimate power in microprocessors," IEEE Transactions on Circuits and Systems II: Express Briefs, 60(12), pp. 882–886, 2013.

[8] K.-J. Lee and K. Skadron, "Using performance counters for runtime temperature sensing in high-performance processors," in the Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium, pp. 8, 2005.

[9] J. Cavazos, G. Fursin, F. Agakov, E. Bonilla, M. O'Boyle, and O. Temam, "Rapidly selecting good compiler optimizations using performance counters," in the proceedings of the International Symposium on Code Generation and Optimization, pp. 185–197, 2007.

[10] F. Gunther and S. Fritsch, "neuralnet: Training of neural networks," The R Journal, 2(1), pp. 30-38, 2010.

Evaluation of Heuristic Algorithms for Solving a Transportation Problem

Kacper Rychard, Iwona Pozniak-Koszalka,
Leszek Koszalka, and Andrzej Kasprzak
Dept. of Systems and Computer Networks
Wroclaw University of Technology
Wroclaw, Poland
e-mail: krychard@gmail.com, {iwona.pozniak-koszalka,
leszek.koszalka, andrzej.kasprzak}@pwr.edu.pl

Dawid Zydek
Dept. of Electrical and Computer Engineering
University of Nevada Las Vegas
Las Vegas, U.S.A
e-mail: dawid.zydek@gmail.com

Abstract **-- This paper concerns different approaches to solve a transportation problem. A new idea for solving the formulated problem is developed. Three algorithms, named Highest Cost Method (HCM), Reverse Vogel's Approximation Method (RVAM), and Reverse Russel's Approximation Method (RRAM), have been created. The properties of these algorithms, including the accuracy and the efficiency, are evaluated on the basis of the simulations made using the designed and implemented experimentation system. Moreover, the paper contains the results of the comparison between known algorithms and the proposed algorithms. The comprehensive studies show that the proposed algorithms are more accurate; however, they require more processing time to find the solution.**

Keywords- transportation problem; algorithm; heuristic; cost reduction; experimentation system.

I. INTRODUCTION

The transportation problem is a well-known issue faced by majority of companies. Transportation is usually the main component of the company's logistics budget [1]. Ineffective transport generates unnecessary costs that can lead to wasting large amounts of money in a scale of a whole company. Even the largest companies take sometimes peculiar actions, e.g., avoiding the left turns in routes of their delivery trucks to reduce the total costs of transportation [2]. Essentially, the main problem is how to move goods from group of m sources to n destinations in a way that minimizes the total transportation cost [3]. As the pace of both industrial and economic development was increasing, more and more goods started to be transported. These changes include an increase in the need for transportation, new types of transported goods, and new ways of transporting them. At some point, the task of cost control in such a system has become too difficult to be performed without specialized tools.

Increasing attention in Internet and network services has created new categories of transportation systems in order to determine traffic of many different applications, e.g., P2P multicast or Content Delivery Networks [4]. These new adoptions of the problem define new requirements for approaches used to solve the problem. First of all, the solution often has to be obtained quickly as the time is the key factor affecting the quality of service and user experience. Moreover, for more complex problems with great number of sources and destinations, obtaining the optimal solution is often not possible in a reasonable amount of time. Because of that, heuristic algorithms combining time efficiency and capability of obtaining a close to the optimal solution need to be used. However, the accuracy of these algorithms is usually ensured at the expense of additional calculations. This creates the challenge of balancing between the short processing time and precise calculations when finding a way to solve the transportation problem. This paper is an extension of our work in [5]. We propose three new algorithms:

- Highest Cost Method (HCM),
- Reverse Vogel's Approximation Method (RVAM),
- Reverse Russel's Approximation Method (RRAM).

The rest of the paper is organized as follows. Section II presents the related work. Section III describes the transportation problem's mathematical model and its most popular representations. Section IV includes a description of the main algorithms and the most important pros and cons of their use. A new approach with the proposed algorithms is described in Section V. Section VI and Section VII contain the design of the experiments, next, their results, followed by comments. The developed tool and its use are also presented. The paper is concluded in Section VIII followed by the plans for the future work.

II. RELATED WORK

In [6], a fuzzy version of the transportation problem with additional restrictions is examined. The fuzzy transportation problem is characterized by fuzzy intervals as the unit costs of the shipment links. In this paper, the problem was transformed into the classical linear fractional programming problem presented in [7]. A time minimization in a fuzzy version of the problem is a subject of the research - in [8], the authors present a procedure to obtain an optimal solution which provides the longest time on active transportation routes as well. A numerical example is included to validate the presented approach. The problem with uncertain cost, supplies, and demands rather than fuzzy variables is studied in [9], where the authors discuss the possibility of transformation of the problem into the deterministic form. Problems

ISBN: 978-1-61208-432-9

that take into account both cost and time are presented in [10]. However, this work is focused on finding new ways of solving cost transportation problems - two new methods (blocking method and blocking zero point method) are proposed. While most of the methods solving the transportation problem focus on minimizing only one factor, the work in [11] solves the problem in such a way that both cost and time are minimized. In [12], the novel Artificial Immune Algorithm is presented to solve the Fixed-Charge Transportation Problem. In this modification of the original transportation problem, the total cost of transportation depends on the unit costs and on additional cost associated with the link use. The authors compare their work to most recent methods [13] sowing that the proposed procedure is superior to them. An ant colony optimization algorithm is presented in [14] as well as an approach which uses both genetic algorithm and local search in solving multi-objective transportation problem. The paper considers the problem with a cross-docking network. The proposed algorithms reduce the total cost in some type of transportation networks and perform better than Branch-and-Bound method [15]. The authors emphasized the importance of heuristic algorithms, in particular hybrid evolutionary algorithms in optimization problems.

III. TRANSPORTATION PROBLEM STATEMENT

The main assumptions of the problem are that the cost of transportation between a given source and destination depends on the quantity of goods transported (all the unit costs are known) and the acceptable solution is the one that exhausts supplies of all sources and fulfills demands of all destinations without the negative values of allocations [16]. The considered problem is formulated as a set of formulas:

$$C(X) = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} \to \min \qquad (1)$$

$$\sum_{j=1}^{n} x_{ij} = s_i \qquad (2)$$

$$\sum_{i=1}^{m} x_{ij} = d_j \qquad (3)$$

$$x_{ij} \geq 0 \qquad i = 1,2,...,m \qquad j = 1,2,...,n \qquad (4)$$

$$\sum_{i=1}^{m} s_i = \sum_{j=1}^{n} d_j \qquad (5)$$

The above expressions are described in the following way: The total cost of the problem should be minimal, where *C(X)* is the total cost, c_{ij} are the unit costs, and x_{ij} represent allocations (1). The total amount of goods sent from each source should be equal to its supply, where s_i are the sources' supplies (2). The total amount of goods sent to each destination should be equal to its demand - d_i are the destinations' demands (3). All allocations should be non-negative (4). In the balanced problems, the equation (5) states that the sum of all supplies s_i equals the sum of all demands d_j, which means that there is a solution exhausting all sources' supplies and fulfilling all destinations' demands.

A graph representation of the problem is shown in Fig. 1. The sources and destinations are represented by circles; they are denoted by s_i and d_i stand for the sources' supplies and destinations' demands, respectively. The arrows represent the shipping links and the numbers placed on them are the unit costs [16].

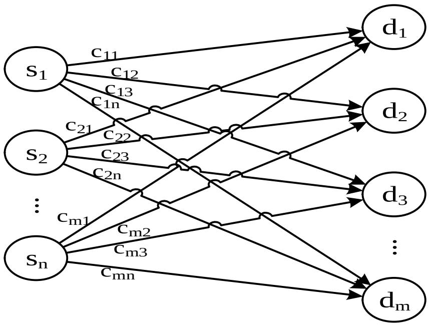


Figure 1. Graph representation of the transportation problem.

The same problem may be illustrated by Table I.

TABLE I. MATRIX REPRESENTATION OF THE PROBLEM

	d_1	d_2	d_3	$\cdots$	d_n
s_1	c_{11}	c_{12}	c_{13}	$\cdots$	c_{1n}
s_2	c_{21}	c_{22}	c_{23}	$\cdots$	c_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_m	c_{m1}	c_{m2}	c_{m3}	$\cdots$	c_{mn}

The table shows the matrix of costs (c_{ij}) and two vectors representing the supplies of the sources (s_i) and the demands of the destinations *(d_j)*.

IV. KNOWN ALGORITHMS

The most common algorithms for solving transportation problems are based on the triangularity rule [3] saying that the feasible solution is obtained after the operations:

Step 1. If the problem consists of only one source or destination, set all the amounts of transportation to highest possible. Go to STOP.

Step 2. For the next link *xij* to be considered, set the amount of transportation to highest possible. Reduce the problem using *i*-th source or *j*-th destination.

Step 3. Go to step 1.

The maximum possible amount of transportation in each shipment link is calculated as the smaller number from the two numbers of: supply of the source and demand of the destination linked. It is noteworthy that the main idea of the algorithm remains all the time the same; however, the technique sometimes is described differently. The changes concern a number of steps and, what is more important, the possibility of deleting the source and destination in one step. This action results in obtaining the solution that is degenerated (uses less than $m+n-1$ shipment links). Deleting only one source/destination at the time leads to allocations with the value of 0 in the solution, which does not increase the cost of the solution found, but allows using the solution as an input to optimization algorithm.

North-West Corner Rule (NCR). In this algorithm, the links are considered in a sequence as they appear in the problem matrix. This method is supposed to provide a fast way of achieving a feasible, but not necessarily efficient solution [17]. The step list of the NCR algorithm is:

Step 1. If the problem consists of only one source or destination, set all the amounts of transportation to highest possible. Go to STOP.

Step 2. For the North-West link (top left element in the matrix of costs) x_{ij}:

a. If $s_i > d_j$ (supply higher than demand), then allocate *dj* to this link. Decrease the *si (i*-th source's supply) by d_j. Delete the *j*-th destination.

b. If $s_i < d_j$ (demand higher than supply), then allocate s_i to this link. Decrease the d_j (*j*-th destination's demand) by s_i. Delete *i*-th source.

c. If $s_i = d_j$ (supply equals demand), then choose randomly action of 2a or 2b.

Step 3. Go to step 1.

The main advantage of this approach is that not all of the shipment links are considered by the algorithm. Once the source/destination is deleted, all other links leading from/to node are omitted.

Lowest Cost Method (LCM). The main idea behind LCM is to sort the connections by the unit cost *cij* and use the cheapest ones first [3]. However, it is unlikely to use this method to provide the final solution. The idea of using the LCM as a heuristic algorithm is novel. The main advantages of this approach are the simplicity, quickness, and way better solution than the NCR. The solution returned by this algorithm meets all the main requirements (exhausting all sources' supplies *si* and fulfilling all destinations' demands *dj* without negative amount of transportation) and is supposed to be closer the optimal solution than the output returned by the NCR. The step list of the LCM algorithm is almost identical to the NCR. The only difference is in the order of the links considered in step 3.

Vogel's Approximation Method (VAM). The VAM is in some sense an extension of the LCM. According to this algorithm, the unit cost of the link is not the only determinant of its position in a sequence. More important is the difference between the lowest unit cost *cij* in a row/column in the matrix of costs and the second smallest one. As the next link to be considered, the one with the lowest unit cost *cij* in the chosen row/column is selected [3]. The process of the link selection in the VAM is as follows: (i) For each row and column calculate the difference between two lowest values of the unit costs in this row/column; (ii) Select the row/column with the highest value of calculated difference; (iii) Consider the link with the lowest unit cost in the selected row/column as a next.

Russel's Approximation Method (RAM). The RAM, like the VAM, depends on the calculations made on the matrix of costs while choosing the next link to consider. In each step the maximum costs in each i-th row and j-th column are found. The assist value (γ_{ij}) is calculated to determine the next link. The link with the lowest γ_{ij} value is selected.

Optimization algorithm. This algorithm takes any valid solution of the problem as an input and gives the best possible solution as an output. It checks the optimality of the solution and finds the non-used connection that should be used to reduce the total cost of transportation (if the solution was not optimal). Adding the connection to the solution may increase or decrease other allocations. Next, the described steps are repeated. The algorithm stops when the solution is optimal. The number of iterations done varies and depends on the input solution – mostly on its accuracy. The detailed description of how the algorithm works can be found in [3].

V. THE PROPOSED ALGORITHMS: HCM, RVAM, RRAM.

The main idea to design the proposed algorithms was based on the approach that is opposite to the LCM. If it is possible to use the cheapest links to transport goods, avoiding the use of links with highest unit cost should result in a similar solution.

The main problem of this approach is that the calculations of the minimum allocations assume that the values of the supplies and demands will not change and it will be possible to exhaust current supply/fulfill demand elsewhere in further steps of the algorithm. When some allocations are made, the previous assumptions cease to be valid. This is why the algorithm calls itself with some amounts of transportation pre-allocated and the corresponding supplies and demands decreased [5]. In every step, the minimum allocation is calculated as:

$$x_{ij\,\min} = \max\{s_i - d'_j, d_j - s'_i, 0\}\,, \quad s'_i = \sum_{\substack{k \neq i \\ e_{kj}=0}} s_k\,, \quad d'_j = \sum_{\substack{l \neq j \\ e_{il}=0}} d_l \qquad (6)$$

In (6), the *s'*$_i$ stands for the sum of all the unconsidered supplies in this run of the algorithm except *i*-th; the d'_j stands for all unconsidered demands in this run of the algorithm except *j*-th; the e_{ij} is a variable responsible for determining whether the link between *i*-th source and *j*-th destination was considered in this run of the algorithm (1 if it is true, 0 otherwise). When a given source has enough other destinations to send goods to, and a given destination has enough other sources to receive goods from, both $s_i - d'_j$ and $d_j - s'_i$ are less than 0 and x_{ij} min gets the value 0. To improve algorithm's accuracy, the case when $s_i - d'_j = 0$ or $d_j - s'_i = 0$ is distinguished. When an allocation due to the source is made, the amounts of transportation in all the other unconsidered links of this source are set to their maximum (calculating the minimum in the first link was based on the assumption of allocating the maximum in all the other links). The source and the destinations for which the maximum allocations were made are deleted and the algorithm repeats. If an allocation is caused both by the source and destination, only one of the above action chains is taken. The step list is as follows:

Step 1. If the problem consists of only one source or destination, set all the amounts of transportation to highest possible. Go to STOP.

Step 2. For the next link x_{ij} is to be considered:

(a) If $s_i - d'_j < 0$ and $d_j - s'_i < 0$, then mark link as considered. Go to step 3.

(b) If $s_i - d'_j \geq 0$ or $d_j - s'_i \geq 0$, then choose the case:

(*i*) If $s_i - d'_j > d_j - s'_i$ (allocation due to the source), then allocate $s_i - d'_j$ in this link and maximum in all of the other unconsidered links of this source. Delete *i*-th source and the destinations for which the maximum allocations were made.

(*ii*) If $s_i - d'_j < d_j - s'_i$ (allocation because of the destination), then allocate $dj - s'i$ in this link and maximum in all of the other unconsidered links of this destination. Erase the *j*-th destination and the sources for which the maximum allocations were made.

(*iii*) If $s_i - d'_j = d_j - s'_i$, then choose randomly action from step 2b(*i*) or 2b(*ii*).

Step 3. Go to step 1.

The presented idea may be also used for creating reverse versions of VAM and RAM - every single allocation results with deleting one source or destination before the algorithm continues until the only source or destination is preserved. Then, the remaining allocations are made. The total number of links used in the returned solution equals $m + n - 1$. It is the exact number of the used links in the solutions obtained by the methods based on the triangularity rule.

Highest Cost Method (HCM). This algorithm is a developed version of Expensive Means Less (EML), which was presented by the authors in [5]. The HCM is based on the main proposed idea of avoiding allocations on links with high unit cost. In this algorithm, the most expensive links are considered first. It is supposed to return solutions with cost similar to those of the LCM, but recursive calls may cause increase of its overall runtime. The only factor determining the order of the considered links is their unit cost. This algorithm, as well as the LCM, is characterized by kind of 'shortsightedness' - it does not take into account the consequences of made decisions, e.g., avoiding one expensive link may cause the necessity of the use a few others later one, leading to an increase in the overall cost of the solution.

Reverse Vogel's Approximation Method (RVAM). The RVAM is based on the VAM algorithm, but it uses different priorities during determining the next link to be considered. While the VAM selects the minimum element of the row/column of the cost matrix with the biggest difference between two smallest elements, the RVAM chooses the maximum element of the row/column with the biggest difference between two most expensive links. This can be interpreted as seeking the link which, if avoided, would potentially prevent the cost to increase the most. The step list of the link selection process in the RVAM is as follows: (i) For each row and column calculate the difference between two highest values of the unit costs in this row/column; (ii) Select the row/column with the highest value of the calculated difference; (iii) Select the link with the highest unit cost in the selected row/column; (iv) Consider this link as next.

Reverse Russel's Approximation Method (RRAM). As in the case of the VAM and the RVAM, the RRAM combines the original RAM approach with the proposed idea. To select the next link to be considered, the assist value γij is calculated. The link with minimum γij is to be chosen.

VI. EXPERIMENTS

The objective was to test the efficiency and accuracy of the implemented algorithms. The testing tool is an application implemented in C# language using Microsoft Visual Studio 2010. Class library ZedGraph was used to draw charts and present the effect of the tests in a graphical form.

Experimentation system. The implemented testing tool allows the user to select the range of the input data. The final solution is an average of the tests' results. As for the number of goods transported parameter, the user is allowed to input the average supply and the demand is calculated to balance the problem. As the number of sources (*m*) and destinations (*n*) varies during the tests, the problem of maximum size is generated and on each step of the test it is reduced to proper size. Then, the problem is balanced by increasing the supply of the last source (*sm*) or demand of last destination (*dn*) accordingly.

The tests were designed to deliver the information about the main characteristics of the implemented algorithm, which are processing time and cost found. To allow a more valuable analysis, it is possible to get information about processing time and cost reduction with optimization algorithm enabled. Before the main part of the experiments, the preliminary experiment was made to determine how the results depend on the characteristics of the input data and how to choose the input data to make the tests more reliable. The experimentation system may be regarded as input-output system (the block-diagram is shown in Fig. 2).

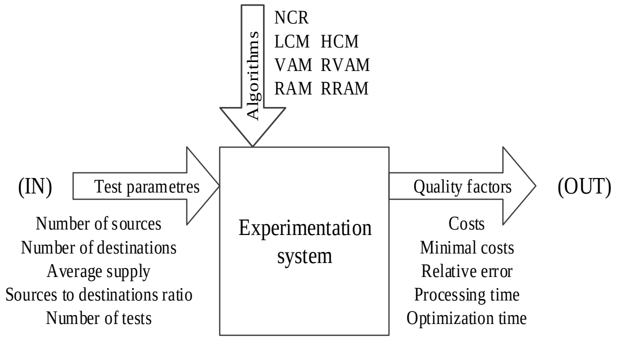


Figure 2. Experimentation system as input-output plant.

Experiments were conducted in order to investigate:

(*i*) Cost of the solution found by algorithms without optimization in comparison to the optimal one (depending on the size of the input);

(*ii*) Relative error of solutions found by algorithms without optimization (depending on the size of the input); (*iii*) Processing time of the algorithms (depending on the size of the input).

All the experiments were made with the number of tests (single experiments) set to 10.

Preliminary experiment. The experiment consisted in testing the total cost expressed by (1) obtained by the algorithms depending on the size of the problem (m x n) defined by the data matrix (see Table 1). For all algorithms, the same matrices (the same size of the problem) were tested, e.g., 4 × 6, 6 × 4, 2 × 12, etc.

For any matrix, single experiments were repeated and the averaged values were treated as the results of the experiments. In Fig. 3, NCR and LCM solutions as well as the optimal cost (marked as the OPT) are presented.

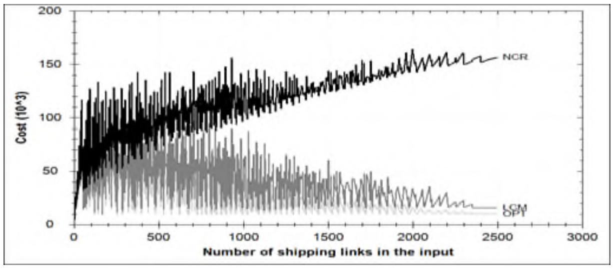


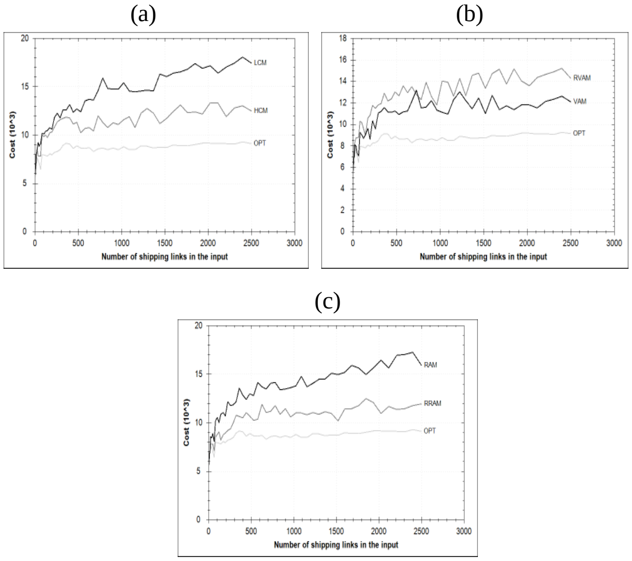
Figure 3. Preliminary experiment results.

It may be observed that NCR algorithm is about eight times less accurate than LCM. Therefore, NCR was excluded from further experiments. In the next experiments, the data in a range from 1 × 1 to 50 × 50 with the same number of sources and destinations *(m = n*) were taken into consideration.

VII. ANALYSIS OF THE RESULTS OF EXPERIMENTS

A. Cost of the Solution

The experiment was designed for finding the relationship between the cost expressed by (1) and produced by the known and proposed algorithms and the size of the input measured by the number of shipping links. The results are shown in Fig. 4.


(a)

(b)

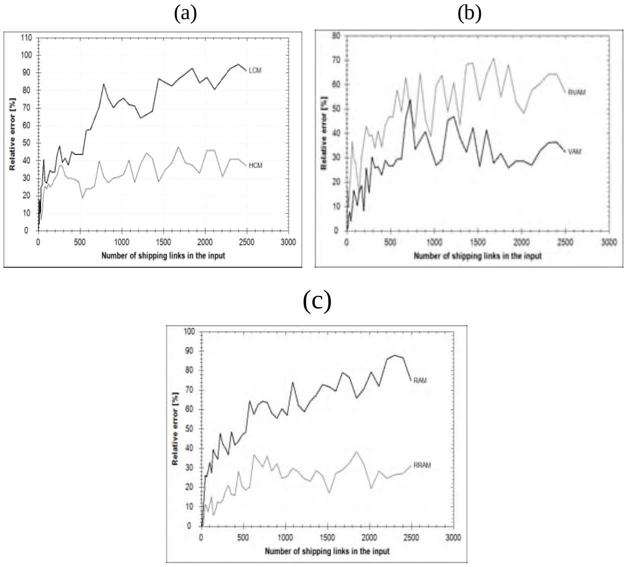
(c)

Figure 4. Cost comparison: (a) LCM with HCM, (b) VAM with RVAM, (c) RAM with RRAM.

Two of the proposed algorithms (the HCM and the RRAM) returned the solutions better than their known equivalents. The HCM outperforms the LCM by approximately 29%. The RRAM returns solutions that are cheaper than the ones returned by the RAM by approximately 25%. In case of the VAM and the RVAM, the proposed algorithm is not as good as the original one. The solutions of the VAM are about 15% cheaper than those found by the RVAM.

B. Relative Error

To determine the accuracy of the algorithms, the relative error of the returned solutions was examined as well. It was calculated as the ratio of the cost of the solution found to the minimal possible one (i.e., the cost of the optimal solution). The results are shown in Fig. 5.


(a)

(b)

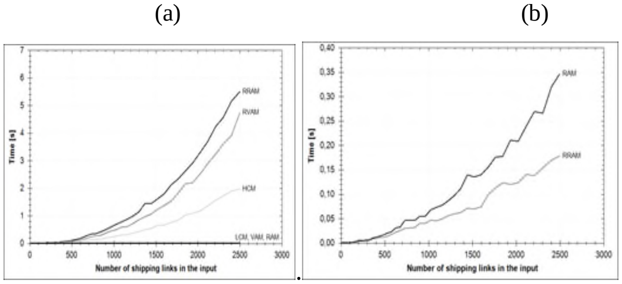
(c)

Figure 5. Relative error comparison: (a) *LCM* with *HCM, (b) VAM* with *RVAM, (c) RAM* with *RRAM.*

It may be observed that HCM and RRAM are found to be remarkably more accurate than LCM and RAM by 43% and 34%, respectively. In opposite, RVAM performed worse than VAM by 24 %. It is worth to notice that the relative error of all tested algorithms was correlated to the size of the problem.

C. Processing Time

Two cases were considered: (i) The time of finding solution without optimization (called as algorithm alone), (ii) The time spent on performing optimization procedure. The results are shown in Fig. 6.


(a)

(b)

Figure 6. Processing time comparison: (a) algorithms alone, (b) algorithms with optimization.

The comparison of the total processing times needed by algorithms for finding the solution with optimization is given in Fig. 7.

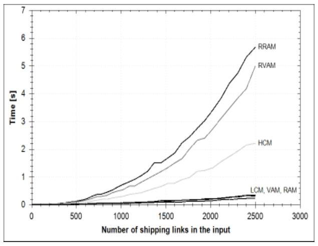


Figure 7. Processing time comparison – the total processing

The created algorithms take more time to calculate the solution. It can be explained by the recursive calls present in the proposed approach. The optimization based on the HCM and the RRAM solutions runs faster than the cases of the LCM and the RAM. However, what was expected, the total processing time is shorter for the classical algorithms.

VIII. CONCLUSION AND FUTURE WORK

In the paper, three created algorithms for solving the transportation problem were presented, evaluated, and compared with the existing ones. The HCM and RRAM found better solutions than their literature equivalents. The solutions closest to the optimal one were obtained by RRAM. However, from the processing time point of view, the proposed approach is more time consuming. Reasons of this fact should be sought in a more complex way of calculating, i.e., in recursive calls. The fastest is LCM, which sorts the shipment links considering their unit cost only.

In general, we may conclude that, if the processing time of the HCM and the RRAM satisfies the requirements of the logistic systems, they can be used to provide a solution more accurate than those of known algorithms.

As the future work, the influence of the ‘shape’ of the problem (number of sources *m* to number of destinations *n* ratio) on the results is to be tested, and extensions of the proposed algorithms to some modifications of the problem, such as fixed charge transportation, will be considered. The most important are: (i) non-balanced problem [18], (ii) costs of storage/shortage, (iii) blockage of a shipment link [19]. Moreover, it is planned to give opportunities for making tests in automatic way [20], and to apply the idea of the multistage experimentation [21].

ACKNOWLEDGMENT

This work was supported by the statutory funds of the Dept. of Systems and Computer Networks, Faculty of Electronics, Wroclaw University of Technology, Poland.

REFERENCES

[1] M. Morrison, “Optimize your transportation program for greater efficiency”, [online] Logitics.com/digital/issues/digital.pdf [retrieved: March, 2015].

[2] J. Lovell, “Left-Hand-Turn Elimination,” article in: The New York Times Magazine, Dec 9, 2007.

[3] G. B. Dantzig and M. N. Thappa, Linear Programming 2: Theory and Extensions, Springer, New York, 2003.

[4] C. Huang, N. Holt, Y. A. Wang, A. Greenberg, J. Li, and K.W. Ross, “A DNS reflection method for global traffic management”, Proc. USENIX Annual Techn. Conference, Boston, 2010, pp. 265-270.

[5] K. Rychard, W. Kmiecik, L. Koszalka, and A. Kasprzak, “Experimentation system for heuristic algorithms to solving transportation problem”, Proc. of 8th Intern. Conf. on Systems (ICN'13), IARIA, Seville, 2013, pp. 27-32.

[6] D.Dutta and A. S. Murthy, “ Fuzzy transportation problem with additional restrictions”, ARPN Journal of Engineering and Applied Sciences, vol. 5, 2010, pp. 36-40.

[7] G. Sharma, S. H. Abbas, and V. K. Gupta, “ Solving transportation problem with the various methods of linear programming problem”, Asian Journal of Current Engineering and Maths., vol. 1, 2012, pp. 81-83.

[8] J. M. Vinotha, W. Ritha, and A. Abraham, “Total time minimization of fuzzy transportation problem”, Journal of Intelligent & Fuzzy Systems, vol. 23, 2012, pp. 93-99.

[9] Y. Sheng and K. Yao, “A transportation model with uncertain costs and demands”, International Interdisciplinary Journal, vol. 15, 2012, pp. 3179-3186.

[10] P. Pandian and G. Natarajan, “A new method for solving bottleneck-cost transportation problems,” International Math. Forum, vol. 6, 2011, pp. 451-460.

[11] S. K. Kumar, I. B. Lai, and S. B. Lai, “Fixed – charge Bi-criterion transportation problem”, International Journal on Computer Applications, vol. 2, 2012, pp. 39-42.

[12] K. M. Altassan, M. M. El-Sherbiny, Y. M. Ibrahim, and A. D. Abid, “A novel artificial immune algorithm for solving fixed charge transportation problems”, Proc. of Conf. on Computer Science and Inform. Technology, vol. 36, 2013, pp. 114-121.

[13] M. Hajiaghaei-Keshtelia, S. Molla-Alizadeh-Zavardehib, and R. Tavakkoli-Moghaddama, “Addressing a nonlinear fixed-charge transportation problem using a spanning tree-based genetic algorithm”, Computers & Industrial Engineering, vol. 55, 2010, pp. 259-271.

[14] R. Musa, J. P. Arnaout, and H. Jung, “Ant colony optimization algorithm to solve the transportation problem of cross-docking network”, Computers & Industrial Engineering, vol. 59, 2010, pp. 85-92.

[15] S. A. Zaki, A. A. Mousa, H. M. Geneedi, and A. Y. Elmekawy, “Efficient multiobjective genetic algorithm for solving transportation, and transshipment problems”, Appl. Math.. vol. 3, 2012, pp. 92-99.

[16] A. Calczynski, J. Sochanska, and W. Szczepankiewicz, Methods of Shipment Rationalization in Trade, /in Polish/, WAE, Cracow, 1998.

[17] F. S. Hillier and G. J. Lieberman, Introduction to Operations Research, McGraw Hill, New York, 2012.

[18] A. Marczuk and W. Misztal, “ Agricultural produce transport optimization in the conditions of market non-balance” /in Polish/, Journal Inz. Rolnicza, vol. 129, 2012, pp. 221-226.

[19] K. Pienkosz, “ Optimization models and methods of resource allocation,” /in Polish/, Scientific Reports Series Elektronics, Warsaw University of Technology, vol. 3, 2010, pp. 124-132.

[20] D. Ohia, L. Koszalka, and A. Kasprzak, “Evolutionary algorithm for solving congestion problem in computer network", Lecture Notes in Computer Science, Springer, vol. 5711, 2009, pp. 112-121.

[21] L. Koszalka, D. Lisowski, and I. Pozniak-Koszalka, “Comparison of allocation algorithms for mesh structured networks using multistage simulation”, Lecture Notes in Computer Science, Springer, vol. 3984, 2006, pp. 58-67.