CERC2016

Collaborative European Research Conference

Cork Institute of Technology – Cork, Ireland

23 - 24 September 2016

www.cerc–conference.eu

Proceedings

22 Extended - Abstracts
and 14 selected Full Papers

Version 2 (January 2017)

Editors :
Udo Bleimann
Bernhard Humm
Robert Loew
Ingo Stengel
Paul Walsh

ISSN : 2220 - 4164

Hints :
- Topics in the table of content are linked to page where that abstract begins.
- The page number at the end of all pages is linked to begin of table of content.

Table of Content

Extended Abstracts and selected Full Papers

Chapter 1
Business

Chapter 2
Business Information Systems

Chapter 3
Applied Computing

Chapter 4
Computing

Chapter 5

Bioinformatics

Chapter 6

Humanities / Social

Call for Papers

List of Authors

Preface

In today's world, which has recently seen fractures and isolation forming among states, international and interdisciplinary collaboration has become an important source of progress. Collaboration is a rich source of innovation and growth and it is the goal of the Collaborative European Research Conference (CERC 2016) to foster collaboration among friends and colleagues across disciplines and nations within Europe. CERC emerged from a long-standing cooperation between the Cork Institute of Technology, Ireland and Darmstadt University of Applied Sciences, Germany. CERC has grown to include more well-established partners in Germany (Hochschule Karlsruhe and FernuniversitÃ¤t Hagen), United Kingdom, Greece, Spain, Italy, and many more.

CERC is truly interdisciplinary, bringing together new and experienced researchers from science, engineering, business, humanities, and the arts. At CERC researchers not only present their findings as published in their research papers. They are also challenged to collaboratively work out joint aspects of their research during conference sessions and informal social events and gatherings.

To organize such an event involves the hard work of a number of people. Thanks go to the international program committee and my fellow program chairs, particularly to Prof Dr Udo Bleimann and Prof Dr Ingo Stengel for organizing the conference and the review process. Dr Robert Loew put a great effort into preparing the conference programme and proceedings. Prof Udo Bleimann was invaluable for local organization. Thanks also to Dr Brendan Murphy and Tim Horgan for supporting CERC.

Dr Paul Walsh

General Conference & Programme Co-Chair, CERC2016

Cork, September 2016

Keynote
Standards or Excellence in Project Management ?

Prof. Dr. Nino Grau
Technische Hochschule Mittelhessen - University of Applied Sciences, Germany
e-mail: Nino.Grau@wi.thm.de

How do standards for project management support project managers in their everyday work? Are they critical success factors or do they hinder project manager`s creativity on the way to excellence in project management?

Let us have a look on standards like ISO 21 500 and on IPMA PEM Project Excellence Modell.

Short CV Prof. Dr. Nino Grau

He holds degrees in computer science (University of Erlangen) and industrial engineering (Technical University Munich) and received his doctorate with a thesis about decision making in teams.

His main area of interest is project and process management. He was a member of the board and deputy president of GPM the German Project Manager Association. GPM is MA (Member Association) of IPMA the International Project Management Association. IPMA is the oldest international umbrella association for project management associations and repre-sents over 60 MAs with over 60 000 individual members Worldwide.

As a Vice President of IPMA Grau started the IPMA activities concerning the development of the new ISO 21500 standard and has been Head of Delegation representing IPMA at ISO. He was responsible for the IPMA “Young Crew” and “Standards and Awards” within IPMA.

As a professor at the Technische Hochschule Mittelhessen University of Applied Sciences in Friedberg he has been responsible for introducing the first postgraduate degree course in project management in Germany, which started in September 2002.

Chapter 1

Business

Suitable Financing in the Context of a High-Tech (Semiconductor) Start-up

Mahbub Akhter, Pleun Maaskant and Angela Wright*
Tyndall National Institute, University College Cork, UCC, Cork Institute of Technology, CIT*
Corresponding author: mahbub.akhter@tyndall.ie

Keywords: Start-up, Entrepreneur, Venture Capital.

This research investigates the financing options available to a "High-Tech Start-up" when a technical person seeks to start a business from innovations emerging from his or her daily work. In seeking to understand the different factors which might influence the choice of a particular type of financial source, a qualitative research approach was considered to be the most appropriate one. As s part of this qualitative approach, a number of interviews with entrepreneurs, CEOs and CTOs have been conducted. Findings from these people with financial experiences are studied, analysed and compared, to select a suitable financing option for a photonics start-up originating from the work inside a research institute.

Considering the high amount of capital required to start a semiconductor start-up, private equity based Venture Capitalist (VC)s are thought to be the most appropriate funding source for such a type of high-tech start-up. This research explores the different aspects of VC funding and the creation of favourable environments to attract more VC investment in the area of semiconductor start-ups. It is found that, the VCs bring much more to the table than merely finance. The role of the VCs in the start-up extends way beyond the traditional financial aspect, with their experience and expertise in different areas, they become deeply involved with the managing and mentoring of the firms they finance. With all their positive influences, the Venture Capitalists contribute significantly to the success of a high-tech start-up. As Filipov (2011) states that some presently highly successful companies such as Apple Computers, Cisco Systems, Genentech, Intel, Microsoft, Netscape, and Sun Microsystems were initially financed by venture capital organisations (Filipov, 2011: 5).

It is concluded that VCs are not the only solution for financing a high-tech start-up throughout all its stages of development. Also, the interaction between entrepreneurs and venture capitalists is not immune to conflicts; both sides need to promote mutual understanding in order to bring success to the "high tech start-up". The role of the support infrastructure and the formation of the proper team in terms of required qualifications and experience has also been studied.

References

Filipov, G., N. 2011, 'Does Venture Capital Contribute to the Success of Startups?' B. Sc. Thesis, Erasmus University, Holland

A Port-Centric Logistic Supply Chain for Container Traffic: The Case for the Port of Cork Company?

Captain Kevin O'Callaghan, Dr Angela Wright
Department of OPD
CIT – Cork Institute of Technology, Ireland
e-mail: angela.wright@cit.ie

Keywords: Port-Centric Logistics, Ports, Cluster Containerisation, Shipping

As a consequence of globalisation, port performance and integrated port logistics have become increasingly important. Ireland as an Island nation depends on ports for the import and export of 99% of its goods. The only way that Ireland can remain competitive within worldwide markets is to have an efficient logistics supply chain. This collaborative research was undertaken in conjunction with the port of Cork and CIT, and seeks to examine the feasibility of introducing a port-centric logistics supply chain for containers in the Port of Cork. A port-centric logistics supply chain is a futuristic logistics process, whereby the port becomes a logistics hub rather than the traditional role as a simple gateway pier for the transfer of cargo. Currently, there is a dearth of academic research literature examining the case for port-centric logistics in Ireland.

A post-positivistic qualitative research methodology is applied in this new research to investigate the feasibility of a port-centric logistics supply chain for containers in the Port of Cork. This approach is considered the best one to address the aims and objectives of the study. The empirical data was gathered by undertaking semi-structured in-depth interviews with port customers, port management, logistics academics, and a senior government official. In total 10 interviews were undertaken.

Positive findings from this research indicate that both industry and the Port of Cork itself support the implementation of a port-centric logistics hub. Another key finding is that there is the potential for a future partnership between the Port of Cork and Cork International Airport, potentially providing a major boost for the Munster area and a significant collaboration for the future of the region. This research has also identified that the implementation of a port-centric logistics hub will reduce the national carbon footprint.

The detailed findings of this research will support the case for the future implementation of a port-centric logistics hub in the Port of Cork and future collaborations.

Galloping Technology: What risks? The Customer Service & Engagement Dynamic.

William Ryan, Dr Angela Wright
Department of OPD
CIT – Cork Institute of Technology, Ireland
e-mail: angela.wright@cit.ie

Keywords: Customer Services, Customer Engagement Strategy, Customer Loyalty, SMEs, Customer Management, Technology

Traditionally, customer service was very 'mechanistic' but, with the development of customer engagement, a transformation occurred; companies recognize the value of interacting with their customers, and customers need and want interaction. This interaction enables organisations to understand their customers, learn more about their needs and what may be required to satisfy them. The aim of this process is to achieve and enable increased customer loyalty. The result is that organisations are engaging with their customers in a timely fashion, and this change has been facilitated by some major advances in technology, namely the advancement of the internet, social media, and the arrival of many and varied new mobile communication channels and platforms.

A modern reality is that internet usage is continuing to expand rapidly, with global internet usage increasing from 29.2% in 2010, to 40.7% in 2014, while usage in Ireland has increased from 69.9% in 2010, to 79.7% in 2014 (World Bank, 2015). This change is also being replicated in the case of mobile phone usage, with three-quarters of the world's inhabitants now having access to a mobile phone. The number of mobile subscriptions in use worldwide, both pre-paid and post-paid, has grown from fewer than 1 billion in 2000 to over 6 billion in 2012, of which nearly 5 billion is in developing countries (World Bank, 2012).

The mobile phone usage figures are proof that the customer landscape has been altered, and with this change comes vast opportunity. Organisations are now embracing these changes, and having effective, timely two-way relationships that were not possible in the past. These new interactions are delivering immense benefits to companies in terms of customer loyalty, product development, and increased sales opportunities. On a more cautious note, these advancements also come with new risks and new challenges for organisations in the form of negatively critical word of mouth. Customers now have the power to highlight poor performance or poor quality to a mass audience through social media, blogs, and forums, with immediate effect. Customers are now becoming more empowered, and are putting the basics of traditional business to the test. Current changes in customer services, and an organisation's ability to successfully employ customer engagement, will largely determine to what degree a company is successful. Consequently, the researchers consider that this is an opportune time to research dynamic field.

This research study is an exploratory examination of customer service in an Irish context. This research will endeavour to evaluate the impact of customer service on corporate strategy and on an organisation's performance in the context of this macro world. The research seeks to determine if technological change and advancements have impacted on customer services and, especially, customer engagement, causing it to evolve, and to examine the effect, if any, these changes have had on businesses. The research explores the concept of 'loyal customers', the availability of 'best

practice' models to aid organisations employing these concepts within the context of customer services, and to evaluate if components of customer loyalty models are transferrable from large organisations to small medium enterprises.

This research on customer services focuses specifically on the concepts and developments of customer engagement, within an Irish perspective. Adopting an interpretivist's paradigm, a qualitative research process was applied involving semi-structured face to face interviews with ten senior experts on customer services and customer engagement, including senior management of global organisations, chief executive officers, experts with worldwide customer services experience, company directors, and individuals from diverse commercial and sporting organisations. These leaders all operate in roles that have direct customer services interaction, and hold positions that have an influence over the strategy and the future direction of customer services within their organizations. Data analysis was carried out using grounded theory, and due consideration was given to ethical, reliability and validity issues. The resultant views, whilst varied, presented seven core themes that will be outlined in full in the paper.

Findings from this research reveal that customer services is having a significant impact on the corporate strategies of companies in today's dynamic technological business environment, and can be the difference between organisational success or failure. Technology has positively impacted customer engagement where it is now 'the lever' to gain closer relationships with customers. This research has identified a direct correlation between the success of a customer services strategy, and the implementation of an effective customer engagement program.

In the absence of a holistic customer engagement model for 'best practice' to assist organisations in implementing a customer engagement strategy, this study has subsequently developed a model to fill this gap in academic literature, and provide organisations with a model that is built in stages, allowing them to progress to 'best in class' status. This research had also identified transferrable components of the new model for practice to small and medium enterprises.

References

World Bank, (2015). "Internet Usage", available at, http://data.worldbank.org/indicator/IT.NET.USER.P2, retrieved January 26, 2016, 08.45.

World Bank, (2012). "Mobile Phone Access Reaches Three Quarters of Planet's Population", available at, http://www.worldbank.org/en/news/press-release/2012/07/17/mobile-phone-access-reaches-three-quarters-planets-population, retrieved January 26, 2016, 10.07.

Chapter 2

Business Information Systems

The Application of Real Time Analytics Technologies for the Predictive Maintenance of Industrial Facilities in Internet of Things (IoT) Environments. A Discussion on Appropriate Research Methodologies

Thomas Rieger †, Stefanie Regier ††, Ingo Stengel ††, Nathan Clarke †
† School of Computing and Mathematics, Plymouth University, United Kingdom
†† Karlsruhe University of Applied Sciences, Germany
e-mail: thomas.rieger@plymouth.ac.uk

Keywords: PhD Thesis, Research Methodology, Predictive Maintenance

Software systems for the analysis of large amounts of data (Big Data Analytics) are applied in more and more areas. The subject of Predictive Maintenance (PdM) using Big Data Analytics systems is present in industrial manufacturing processes, especially in combination with Internet of Things (IoT) environments. Until now, such systems have been using mainly established batch-oriented approaches. New technologies in the area of real time/streaming analytics open up entirely new technological opportunities (Gupta, et al., 2016). These new opportunities might lead to new and even better models in the area of PdM.

The project discussed here deals with remapping existing models for PdM in IoT environments, e.g. condition based or reliability based models, using the above-mentioned new technological opportunities, and with identifying new models and model combinations. New models need to be defined. A comparison of these with established models using big data will outline possible improvements.

The aim of this abstract is to illustrate the discussion that helps to find a suitable research methodology, since a correct scientific approach differs from a practical software developer approach significantly.

The abstract starts with a brief introduction to the subject area and the goals. By adapting the work packages to selected research methodologies it will then illustrate how to find the correct research methodology. The objective of this abstract is to raise the question if one of the methods demonstrated might be valid.

From an IT perspective, industrial automation systems are about managing systems and processes. Communication is focused on process-, production-, and product data. Until now storing the data for later analysis has not been a priority (Schaeffer, 2015).

Due to intelligent devices, more and more data is generated in industrial systems. Connecting these devices in IoT environments makes it possible to store and analyse this data. As a consequence the use of Big Data Analytics for storage and analysis makes sense. “Internet of Things (IoT) will comprise billions of devices that can sense, communicate, compute and potentially actuate. Data streams coming from these devices will challenge the traditional approaches to data management and contribute to the emerging paradigm of big data.” (Zaslavsky, et al., 2013).

A relevant field of application for the analysis of such data is Predictive Maintenance (PdM). The main purpose of PdM is to anticipate unscheduled outages of machines and plants. Maintenance

costs have the potential to be reduced. The availability of production capacities will be more predictable (Eoda, 2014).

New technologies in the area of real time/streaming analytics make it possible to process large amounts of data in real time. Immediate results, tendencies and trends are thus available. Continuous data streams can be permanently analysed.

Research methodology

After the topic of the research project was clear the following questions are about the research design and how to find the correct research methodology. Formulating a precise research question is an important first step. According to Shuttleworth (2015) a general research question will usually be based around ‘why’ or ‘how’ a certain phenomenon is happening.

The project discussed here deals with new technologies in the area of real time/streaming analytics. The focus lies on PdM in industrial processes. Large amounts of data which are generated at high frequencies are necessary for the effective use of Big Data Technologies. These amounts of data can be found in the industrial sector in IoT environments (Zaslavsky, et al., 2013). This leads to the research question “How can new technological opportunities in the area of real time/streaming analytics improve Big Data Applications for PdM in industrial IoT environments?”

When the research question is defined, the following question arises: How to answer the research question in a scientifically sound approach? Following the relevant literature scientific-theoretical approaches of quantitative, qualitative and mixed studies can be found. Applying the classification profile for study designs defined in literature (Döring, et al., 2016) two potential alternatives in the cases of the scientific-theoretical approach and the study groups can be considered (Figure 1, alternatives are on a grey background).

Qualitative or quantitative Approach

Table 1 gives an overview of the differences between qualitative and quantitative approaches (Xavier, 2012). It seems that the quantitative approach fits better to the described research topic than the qualitative approach.

To carry out a quantitative research approach research hypotheses have to be defined. An extensive analysis needs to be carried out using structured data collection methods. This will help to investigate the defined hypothesis. The collected quantitative (numerical) data needs to be analysed statistically (Döring, et al., 2016).

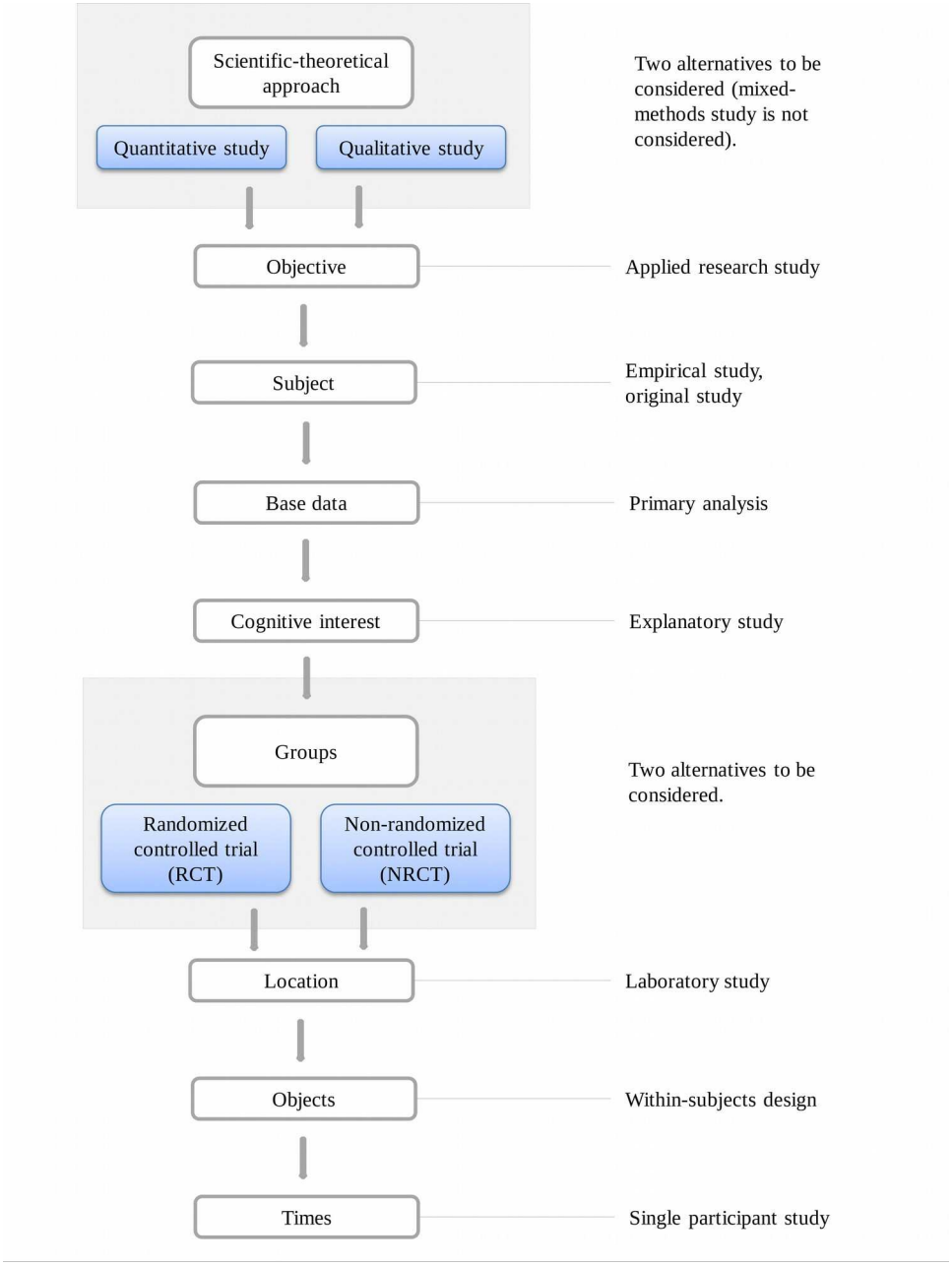


Figure 1: Study design approach according to (Döring, et al., 2016)

Criteria	Qualitative Research	Quantitative Research
Purpose	To understand & interpret social interactions.	To test hypotheses, look at cause & effect, & make predictions.
Group Studied	Smaller & not randomly selected.	Larger & randomly selected.
Variables	Study of the whole, not variables.	Specific variables studied
Type of Data Collected	Words, images, or objects.	Numbers and statistics.
Form of Data Collected	Qualitative data such as open- ended responses, interviews, participant observations, field notes, & reflections.	Quantitative data based on precise measurements using structured & validated data-collection instruments.
Type of Data Analysis	Identify patterns, features, themes.	Identify statistical relationships.
Objectivity and Subjectivity	Subjectivity is expected.	Objectivity is critical.
Role of Researcher	Researcher & their biases may be known to participants in the study, & participant characteristics may be known to the researcher.	Researcher & their biases are not known to participants in the study, & participant characteristics are deliberately hidden from the researcher (double blind studies).
Results	Particular or specialized findings that is less generalizable.	Generalizable findings that can be applied to other populations.
Scientific Method	Exploratory or bottom–up: the researcher generates a new hypothesis and theory from the data collected.	Confirmatory or top-down: the researcher tests the hypothesis and theory with the data.
View of Human Behavior	Dynamic, situational, social, & personal.	Regular & predictable.
Most Common Research Objectives	Explore, discover, & construct.	Describe, explain, & predict.
Focus	Wide-angle lens; examines the breadth & depth of phenomena.	Narrow-angle lens; tests a specific hypotheses.
Nature of Observation	Study behavior in a natural environment.	Study behavior under controlled conditions; isolate causal effects.
Nature of Reality	Multiple realities; subjective.	Single reality; objective.
Final Report	Narrative report with contextual description & direct quotations from research participants.	Statistical report with correlations, comparisons of means, & statistical significance of findings.

Table 1 - QUALITATIVE VERSUS QUANTITATIVE RESEARCH (Xavier, 2012)

An example for a hypothesis could be "This engine is going to fail within a time frame of [X] hours" or "This in-service engine will last [X] more days before it fails" or (comparing batch-oriented PdM to real-time PdM). "The downtime of a machine is reduced from [Y] to [X] hours using the same PdM-Modell/Algorithm with real-time approaches instead of batch-oriented approaches". It is thus necessary to make possible improvements measurable. But, how could valid measurability be achieved?

Industrial environments are diverse and have very different characteristics (e.g. heavy industry vs. pharmaceutical industry). The same is true for IoT systems in these industrial sectors. In order to perform valid measurements and evaluations, a precise application case and a precise environment have to be defined. Application cases in this project use different PdM models (e.g condition-based, model-based, reliability-centred, multi-threshold or continuous-time, mathematical models or empirical approaches). The decision was made to focus on a few selected models, such as remaining lifetime using sensor-based degradation models (Kaiser, et al., 2009).

As the project is mainly data-driven, the criterion for the selection of an environment is the availability of valid measurement data, including as much historical data as possible. There is a choice between generated data, open data repositories or real life data. Current research outlines that open data repositories like the Turbofan Engine Degradation Simulation Data Set (Saxena, et al., 2008) are a good choice because they normally contain comprehensive metadata in addition to existing massive outcome data. The metadata often includes occurred error events and additional information for validation, labelling and classification. The data sets can be used for development of prognostic algorithms (Saxena, et al., 2008). Just like real life data, data from an open data repository is only valid for a specific case of application. Thus, the whole project must be focused

on this application case exclusively. The title of the project and the research question have to be extended by the specific application case.

Once the application case and the environment have been defined, the question of measurability remains. For this purpose Key Performance Indicators (KPIs) will be applied. Even technological objectives, such as "faster availability of results" or "reduction of unscheduled outages" can be defined as KPIs. Ways to identify KPIs include the categorisation via success factors (Parmenter, 2015). Success factors could be, for example, the cost savings in EUR, or the reduction of unproductive time in hours. The following table (Table 2) shows an example for this:

CHARACTERISTIC	UNIT	MEASURE	TYPE
Maintenance cost	EURO	Reduction in %	quantitative
Machine damage	EURO	Reduction in %	quantitative
Spare part storage	EURO	Reduction in %	quantitative
Machine outage	Time	Reduction in %	quantitative
Overtime expenditure	EURO	Reduction in %	quantitative
Machine service life	Time	Increase in %	quantitative
Profits	EURO	Increase in %	quantitative

Table 2 - Success factors according to (Eoda, 2014)

Defining groups

The second open question according to Figure 1 is about the creation and treatment of analysed groups. There are three alternative approaches to be considered: randomized controlled trial (RCT), non-randomized controlled trial (NRCT), and non-experimental study (Döring, et al., 2016).

Non-experimental studies (or descriptive studies) only cover found variations and no experimental variations of independent variables. Non-experimental studies are only very conditionally suitable for the evaluation of a causality hypothesis (Döring, et al., 2016).

A real experiment occurs if randomization or random assignment of groups is performed. An experiment without randomization is referred as a quasi-experiment. If and how groups can be defined for the described project is a topic for discussions (Döring, et al., 2016).

Conclusions

Once the subject of a research project has been found, it is important to define the goal, the major research questions and the approach. This is indispensable in order to ensure a scientifically sound approach and the generating of valid results. The approach already has to be defined at the beginning of the project in order to avoid wrong turns in the course of work.

This abstract describes the journey towards finding a suitable approach for the research project. It has been demonstrated which approaches have been taken for this purpose and which issues still need to be clarified.

References

S. Gupta; S. Saxena, Packt Publishing Ltd. (2016). Realtime Big Data Analytics, First Edition February 2016

M. Shuttleworth, Explorable.com (2015), How to write a Research Paper, 2015

A. Zaslavsky, C. Perera, D. Georgakopoulos (Jan 2013), Sensing as a Service and Big Data, URL: https://arxiv.org/abs/1301.0159, last accessed 08/08/2016

K. A. Kaiser, N. Z. Gebraeel, Predictive Maintenance Management Using Sensor-Based Degradation Models, IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans, August 2009

Eoda GmbH, Whitepaper Predictive Maintenance (mit R), eoda, Februar 2014,
URL: https://www.eoda.de/files/Use_Case_Seiten/Whitepaper/Predictive_Maintenance_mit_R.pdf , last accessed 08/08/2016

N. Döring, J. Bortz, Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer, 5. Auflage 2016

A. Saxena and K. Goebel (2008). Turbofan Engine Degradation Simulation Data Set, NASA Ames Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA, URL: http://ti.arc.nasa.gov/project/prognostic-data-repository, last accessed 08/08/2016

C. Schaeffer, Big Data + The Internet of Things = Big Manufacturing Opportunity. Vantive Media CEO & Blogger, 2015 URL: http://www.crmsearch.com/internetofthings.php, last accessed 08/08/2016

D. Parmenter, Key Performance Indicators: Developing, Implementing, and Using Winning KPIs, John Wiley & Sons, Third Edition 2015

Xavier Unversitiy Library, QUALITATIVE VERSUS QUANTITATIVE RESEARCH, 2012
URL: http://www.xavier.edu/library/students/documents/qualitative_quantitative.pdf, last accessed 08/08/2016

Practical Relevance and Operational Scenarios of Semantic Wiki

Matthias Coym, Philipp Ertelt, Udo Bleimann
University of Applied Science Darmstadt, Department of Computer Science, Germany
e-mail: matthias.coym@h-da.de, philipp.ertelt@h-da.de

Keywords: knowledge, semantic, wiki

Introduction

To keep up with the top contestants on the market, knowledge management is a crucial topic for today's companies. Every organizations resolution is to improve the way it does business. Achieving this goal is strongly dependent on experience and knowledge of the company and its employees.

One approach of knowledge management in a business is the deployment of a wiki. A wiki is an online encyclopedia with articles or entries which can be accessed by the employees. The articles can provide information about different procedures or workflows in the company. Futhermore, it can contain information about customers, projects and experiences. A main fact of a wiki is that it is mostly used as an open encyclopedia. That means employees may be able to not just read, but create new and edit existing articles. This is an agile way to improve and extend the platform. It can be used to concentrate the knowledge in just one system to reduce redundancy and to avoid outdated information, always supposing that the platform is maintained regularly. New employees can use the existing information to have a better start, in contrast, the company keeps the knowledge of employees who leave the company if it is documented it in the wiki.

A wiki does not have to be set in a business case, but can also be used to make information about a certain area accessible to anyone. The most popular example is the free internet encyclopedia platform Wikipedia[1].

A way to extend a wiki platform is the usage of semantic. Semantic describes the relationship between words or phrases in a context. A subject can be connected to an object using a predicate, e.g., university has students. The idea to connect subjects and objects is to give information about relations which are maybe not visible on first sight. This information is called metadata and contain information about characteristics. A semantic network is a representation model of semantic. Figure 1 shows a semantic network where subjects and objects are linked. Each link has a description to define the relation between subjects and objects.

In a semantic wiki the relationships between articles can be filled with information about the data, i.e., it has an underlying knowledge model. The combination of semantic and a wiki can be valuable if used in the right way. This work is about the practical relevance and in which operational scenarios a semantic wiki is useful.

1 www.wikipedia.org

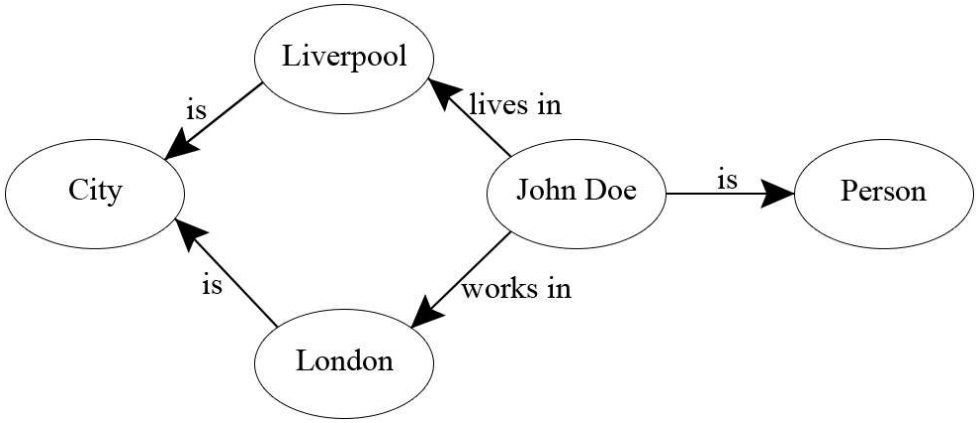


***Figure 1**: Example of a simple semantic network*

Related Research

Several semantic wikis have been developed in recent years. The semantic wiki systems can have very different purposes and functions like personal knowledge management. One of the first semantic wiki systems is the PlatypusWiki, an enhanced Wiki Wiki Web platform (Tazzoli et al., 2004). It uses the Resource Description Framework (RDF), a standard for metadata, and Web Ontology Language (OWL) vocabulary for knowledge representation. It is implemented in Java following the Model View Controller pattern. The focus of the Wiki is set to the creation of RDF data.

Völkel et al. developed an extension for the MediaWiki system called Semantic MediaWiki (Völkel et al., 2006). The extension provides the ability to semantically annotate articles in a wiki. It has also shown that the performance of a semantic wiki has to be considered, since a wiki can become slower with the amount of metadata in the wiki.

Salzburg Research developed a Semantic Wiki called IkeWiki (Schaffert, 2006). IkeWiki has been primarily developed as software for ontology engineering. The wiki data is stored in a Postgres database. It contains two types of content editors. A WYSIWYG editor for novice users as well as a traditional text editor.

Since the Web 2.0 is spread by blogs, wikis, image/audio/video sharing websites, the user behaviour has changed dramatically. Users want to share information and participate among other web users. Therefore, tools are necessary to help users finding the information they search for. Wikipedia has proven to be successful for information interchange. To extend the functionality of a Wiki, the concept of a semantic wiki was born, mixing the features of a wiki and the technologies of a semantic web. Buffa et al. have published in (Buffa, 2008) the SweetWiki. A semantic wiki that combines the advantages of a wiki with the semantic web. Buffa et al. have shown that the use of semantic web concepts can improve navigation, search and usability of a wiki.

In 2005, Krötzsch et al. present a technique to allow knowledge processing in a computer assisted way for Wikipedia (Krötzsch, 2005). The technique called typed links is a simple method for rendering large parts of the Wikipedia platform, so that it is machine readable. Using the technique the user can query intelligently the knowledge database and find related topics of the searched keyword. The method also impacts hardly usability and performance of the wiki.

As aforementioned, a wiki can have different purposes and functions. Oren et al. introduce the prototype of SemperWiki in (Oren, 2006). The wiki is especially for personal and organisational knowledge management. Traditional tools for personal knowledge managements like todo lists or files are very common. But these analogue methods are not automated and cannot be searched digitally. Hence, a digital tool with the requirements to save information, but also provide support finding and reminding information is needed. The advantages of wikis are that they are simple to use and collaborative access is easy to manage, but provide only very limited support for finding and reminding. Oren et al. show how a semantic wiki using metadata annotations improve information access by structured navigation and queries.

Another approach by Tobias Kuhn is the AceWiki (Kuhn, 2008). AceWiki is a prototype of a semantic wiki using the Attempto Controlled English (ACE) for representing its content. ACE is a subset of the English language with limited grammar and fixed formal semantics, i.e., ACE looks like English but has a very restricted syntax. The focus of AceWiki is to improve knowledge aggregation and information representation. The main difference compared to other semantic wikis is, it does not use annotations or metadata for formal statements, but they are the main content of the wiki. An evaluation shows that AceWiki is easy to learn and content can be added quickly by following the three design principles: naturalness, uniformity, and strict user guidance.

KawaWiki is a semantic wiki based on RDF templates and was introduced by Kawamoto et al. in 2006 (Kawamoto, 2006). Describing that for end users it is not easy to create a semantic wiki page to share information from scratch without knowledge of the RDF/OWL syntax, the KawaWiki provides RDF templates to support the author process. The RDF templates used by KawaWiki validate the consistency of the wiki and the created pages. The wiki can generate forms from templates to support the user creating a page and filling it with information. The developed RDF template system by Kawamoto et al. assists expert and non-expert users with a validation checking mechanism to guarantee that the consistency of the wiki is always given.

Methodology

To show how a semantic wiki might be beneficial used, we first tried to understand why semantic wikis are not often used. To do this we looked at the technological aspect of semantic wiki as well as the impact of semantic on the usage and maintenance of the wiki.

Practical Relevance and Operational Scenarios

The practical usage of a wiki system is depending on its purpose. A wiki can be used by a single person as personal knowledge management up to a knowledge management system of a big company to concentrate experience and knowledge in one system, or even a public system like Wikipedia accessible by anyone. A wiki system is mostly used to collaboratively collect and share information to a certain group of users.

A semantic wiki has the benefit over a normal wiki that the underlying semantic net is machine comprehensible. The added meta data enhances the search functionality and can be queried like a database to make use of otherwise unusable connections to automatically generate diagrams and lists.

There are several drawbacks to the use of semantic wiki as well.

There is a conflict between the collaboration aspect and “free for all” approach of a wiki and the care and focus needed for a semantic web. If each writer freely adds relations between pages, the resulting semantic net is supposable complex and contains redundancy. To avoid this a kind of higher administration and/or restriction needs to be set in place, which supposable reduces the motivation to add content or use the semantic features. The addition and maintenance of meta data is extra work for the writer and the use of queries to generate content automatically is not trivial. It is furthermore not enough to just add meta data and relations. A benefit is only achieved, if the semantic data can then be used for non-trivial concrete use cases. This means that not every possible relation is also beneficial to add and advance planning is needed.

Our assumption is therefore that a semantic wiki works best with a wiki of following criteria:

- The content of the wiki is about one specific topic
- Wiki pages correlate to specific objects
- There are multiple cases of the same semantic relations

Conclusion and Future Work

The effort to set up a semantic wiki is reasonable there are also several plug-ins to enhance an existing wiki with semantic. The extra data stored in a wiki can affect the runtime of the wiki, but overall the technical barrier can still be considered low.

A real barrier lies in the conception of beneficial use cases for the semantic data, in the overhead of maintenance for the writers and in the need of a global administration.

The next planned step is to verify our assumption and develop an approach to enhance an existing wiki with semantics based on a concrete use case.

References

Tazzoli, R., Castagna, P., Campanini, S.E.: Towards a Semantic WikiWikiWeb.In: 3rd International Semantic Web Conference (ISWC2004), Hiroshima, Japan (2004)

Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. (2006). Semantic Wikipedia. In Proceedings of the 15th international conference on World Wide Web (WWW '06). ACM, New York, NY, USA, 585-594.

Schaffert, S.: IkeWiki: A Semantic Wiki for Collaborative Knowledge Management.(2006)

Buffa, Michel, et al. "SweetWiki: A semantic wiki." *Web Semantics: Science, Services and Agents on the World Wide Web* 6.1 (2008): 84-97.

Krötzsch, Markus, Denny Vr Denny Vrandecic, and Max Völkel. "Wikipedia and the semantic web-the missing links." *Proceedings of Wikimania 2005*. 2005.

Oren, Eyal, et al. Semantic wikis for personal knowledge management. In: *International Conference on Database and Expert Systems Applications*. Springer Berlin Heidelberg, 2006. S. 509-518.

Kuhn, Tobias. Acewiki: A natural and expressive semantic wiki. *arXiv preprint arXiv:0807.4618*, 2008.

Kawamoto, Kensaku, Yasuhiko Kitamura, and Yuri Tijerino. "Kawawiki: A semantic wiki based on rdf templates." *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*. IEEE, 2006.

Industry 4.0 national and international

Nils Ehlers, Robert Loew, Udo Bleimann
University of Applied Sciences Darmstadt, Department of Computer Science, Germany
e-mail: nils.ehlers@h-da.de, robert@loew.com

Keywords: Knowledge, Industry 4.0, Internet of Things, Industrial Internet Consortium

At the beginning we will explain the term ”Industry 4.0” and the German point of view. In a next step the differences between several countries will be analyzed in more detail.

Germany

International compared Germany is a high-wage region. This leads to the key question if it is possible to maintain and to enhance a production site given the international competition. In contrast to many other industrial countries, Germany was still able to maintain its number of employees in the production sector. The middle class industry plays an important role in Germany. In addition, the small and medium enterprises are the most innovative industries in Germany and helped therefore to minimize the impact of the financial crisis for the German economy, compared to many other countries. The development and the integration of new technologies and processes played a major role in that case. “Industry 4.0” describes the fourth industrial (r)evolution, after the steam machine in 1712, the product line in 1913 and the digital revolution in 1969. Since it will play such an important role for Germany as well as for the rest of the world, it is important for Germany as a production country to maintain and enhance its position.

The development of mechanism with water and steam is called the first industrial revolution. This revolution led to a strong increase in the development of technology, productivity and science.

The second industrial revolution was the development of the production line. Besides new research and knowledge orientated industries, industries in sectors like electrical engineering, chemical industry, mechanical engineering and optical industry took place in Germany. This was also the time when Germany became the leading country in the mechanical engineering industry. This position helped Germany to develop and to use the production line to increase efficiency and to reduce headcount.

The third revolution, also called the digital revolution, describes the digitalization and the related change in industry as well as in many other areas of life the end of the 20th century. Communication played a major role at that time and increased steadily. In 1986 already 20% of the communication was digitalized, 1993 two third and in 2000 already 98% of the whole communication in the world. The digitalization of information processes as well as communication processes led to an information explosion. The global telecommunication capacity as well as the information memory capacity per head increased between 1986 and 2007 about 23-28% per year. These are impressing numbers, especially in comparison with global economy growth of around 3-6% in the same time period.

The fourth revolution is characterized by the interaction between the industrial production with modern information and communication technologies. The German definition of the fourth revolution, based on the “Bundesministerium für Bildung und Forschung”, is: “The future project

Industry 4.0 aims to make the German industry ready for the production in the future." One characteristic is the high level of individualization of the products with the precondition of flexible production for huge amounts of products. Known as a smart factory, a vision of a production environment, production sites as well as logistical systems interact almost fully without the intervention of humans. As a technical basis, cyber-physical systems (CPS) will support the vision under the term "Internet of Things" (IoT). CPS will also allow to produce smart products which have their own memory with all of their characteristics, production processes and determinations. In addition, smart products are also able to exchange that information between other smart products as well as with machines. But not only smart products are able to communicate with each other; also machines will be able to exchange information. This communication is called M2M communication and allows machines to make their own decisions during the whole production process. This communication technology is essential for the effective use of Industry 4.0 and is part of the Internet of Things (IoT) and the Internet of Services. Both terms combined are called the Internet of Things and Services. The relation between IoT and CPS is neither completely clarified nor clearly separated because both concepts were researched and developed simultaneously but they have always been closely related. The term IoT is widely used in America whereas CPS is more common in Europe. Nevertheless, the goal of the total connection und communication of everything is the same for both technologies. Also customers as well as business partners are directly involved in the business processes and the value added processes. The production is also linked with high quality service. With intelligent monitoring and decision processes, companies as well as whole value added processes should be controlled and optimized in real time. Now is the time for Germany to directly shape the fourth industrial revolution. Nevertheless, it is important to keep Industry 4.0 as a research project since it is only depicting a guideline.

Europe

One could assume that under the umbrella of the European Union a centralized guideline for the implementation of Industry 4.0 exists. There are several programs on European level, nevertheless do many European countries have their own initiatives with own goals and funding programs. Besides the German Industry 4.0 project, the EU as well as several countries do have some activities related to the production of the future. On European level a high impact initiative has been started by the European Institute of Innovation & Technology (EIT) with the name "Industry 4.0: Powering Europe". This initiative is under German lead and has the goal to rollout cyber physical platforms in smart factories all over Europe. In addition there is the public-private partnership "Factories of the Future" that focusses on the support of small and medium enterprises enabling success on the global market. Several other countries have also their own activities: Finland wants to spend for his "Industrial Internet Program" around 100 million Euro till 2019. Austria wants to invest 250 million Euro for his Industry 4.0 program. Also France, Great Britain and the Netherlands have started with their own industry 4.0 programs.

USA

In March 2014 the Industrial Internet Consortium (IIC) was founded in the USA. More than 100 companies are participating, like General Electric, IBM, Intel, AT&T and Cisco, to work on the technology for the production of the future. Also important German companies are participating at the consortium, like Siemens and Bosch. Since many IT companies are included, the main focus is on IoT and the interconnection of future factories. The vision of IIC is more or less similar with the goal of Industry 4.0. Production should be more efficient and value added processes should be

optimized. Furthermore, the companies want a higher availability of the machines as well as a high level of individualized production.

China

The Chinese government wants to build up intelligent production systems on the level of omnipresent information systems and communication systems. This is visualized in their "Advanced Manufacturing Technology Roadmap" which is depicting every step till 2050, including milestones for 2020 and 2030. The aim of China is to get less dependend on imported products with the help of technologies for the production, for instance in the automotive sector. This is also clearly visible in the five-year plan since "Advanced Manufacturing Equipment" is one of the areas with the highest priority.

In other countries, like Japan and South Korea, activities are still not clearly known. The visit of many delegations in Germany shows that the interest in this area is extremely high.

References

Toni Pierenkemper: Wirtschaftsgeschichte. Die Entstehung der modernen Volkswirtschaft. Berlin, 2009, S. 88

„The World's Technological Capacity to Store, Communicate, and Compute Information", Martin Hilbert and Priscila López (2011), Science, 332(6025), 60–65

„Videoanimation über The World's Technological Capacity to Store, Communicate, and Compute Information from 1986 to 2010"

Bernd Overmaat, World Wide Wettlauf - Industrie 4.0 in Europa/Asien/USA, https://engineered.thyssenkrupp.com/world-wide-wettlauf-industrie-4-0-aktivitaeten-in-europe-asien-usa/

Was ist Industrie 4.0? In: www.plattform-i40.de. Accessed 9. April 2016.

Chapter 3

Applied Computing

A Clinical Decision Support System for Personalized Medicine

Johannes Idelhauser [1], Bernhard G. Humm [1], Ulrich Beez [1], Paul Walsh [2]

[1] Department of Computer Science
Hochschule Darmstadt – University of Applied Sciences, Darmstadt, Germany

[2] NSilico Lifescience Ltd.
Bishopstown, Co. Cork, Ireland

johannes.idelhauser@stud.h-da.de, {bernhard.humm, ulrich.beez}@h-da.de,
paul.walsh@nsilico.com

Keywords: Clinical Decision Support System, Electronic Health Record, Information Retrieval

Abstract: Given the rapidly growing number of medical publications and resources, consultants face challenges in keeping up-to-date with current research and patient care best practices. This paper presents the concept and prototypical implementation of a Clinical Decision Support System (CDSS) for personalized medicine. It satisfies information needs of consultants at the point of care by integrating secondary medical resources based on a concrete patient's Electronic Health Record (EHR). Particular focus has been on the usability of the CDSS allowing consultants to quickly and intuitively gather relevant information with minimal system interaction. An initial assessment of the CDSS by medical professionals indicate its benefit.

1 Introduction

Given the rapidly growing number of medical publications and resources, consultants face challenges in keeping up-to-date with current research and patient care best practices. Ongoing research and new treatment options require practitioners to keep up-to-date to ensure good treatment outcomes and prevent malpractice lawsuits (Marchant and Lindor, 2013). Physicians' information needs at the point of care range from refreshing and confirming knowledge over logistical questions like drug dosage to teaching and personal learning (Maggio et al., 2014).

Personalized medicine aims at tailoring medical decisions, practices, interventions or products to the individual patient based on their predicted response or risk of disease (Academy of Medical Sciences, 2015). While tailoring the treatment to individual patients is common practise in medicine, the term has been recently used for informatics approaches in medicine that use large amounts of patient data, particularly genomics data, for selecting appropriate therapies.

This paper presents the concept and prototypical implementation of a Clinical Decision Support System (CDSS) (Kawamoto et al., 2005) for personalized medicine. It satisfies information needs of consultants at the point of care by aggregating and integrating primary and secondary medical resources based on a concrete patient's Electronic Health Record (EHR), thus paving the way for personalized medicine. The CDSS is intended to be integrated into an EHR application to be used at the point of care.

The remainder of this paper is structured as follows. Section 2 specifies the problem statement in terms of requirements. Section 3 is the core of the paper describing the concept and prototypical

implementation of the CDSS. Section 4 evaluates our approach. Section 5 compares our approach with related work. Section 6 concludes the paper and indicates future work.

2 Problem Statement

Having consulted extensively with clinicians involved in the treatment of melanoma, we have identified the following requirements:

1. Functional Requirements
    1. Relevant: The CDSS shall satisfy the consultants' information demand with relevant, helpful and latest medical information.
    2. Personalized: The information provided shall be tailored to the medical condition of a particular patient.
    3. Pro-active: The CDSS shall offer information pro-actively without additional data entry by the user.
    4. Easily comprehensible: shall provide a quick overview of all information available as well as the possibility to easily acquire more detailed information where needed.
    5. Workflow: The CDSS shall not interfere with the consultant's EHR workflow.
2. Non-Functional Requirements
    1. Usable: The CDSS shall be intuitive to use and self-explanatory.
    2. Low response time: The response time for all interactions with the CDSS shall be less than 1s.
    3. Extensible: The ongoing extension of the CDSS with new information sources shall be facilitated with moderate implementation effort.

3 A Clinical Decision Support System for Personalized Medicine

3.1 User Interaction Model

Physicians' information needs at the point of care include refreshing and confirming knowledge and logistical questions, e.g., medication dosage, idea generation and personal learning (Maggio et al., 2014). To satisfy these needs, physicians tend to access primary literature in the form of abstracts and full-text as well as secondary literature in the form of summaries and reviews (Maggio et al., 2013). In order to provide an intuitive way for physicians and other health professionals to access this information, the proposed CDSS leverages several information services that each try to satisfy different information needs. Those information services are organized in web page panels that the users can customize and fit to their needs by deciding which service panels should be displayed and which should be hidden. Additionally the order and size of the panels can be adapted to the user's individual needs while the resulting layout is persisted over different sessions for the individual user.

3.1.1 Literature Service

One of the main services in the CDSS is the literature service to find and display relevant primary medical literature that is related to the patient at hand (Figure 1).

Figure 1. Interaction concept to display related medical publications (literature service)

The literature service displays automatically generated filters to quickly navigate the literature search results. The filters are displayed on the left whereas the medical literature is shown on the right. For each medical publication its title, journal and publication date is displayed. In the context of evidence-based medicine (EBM), publications with a high degree of evidence are to be prefered in patient care (Hung et al., 2015). As such, publications that are reviews or clinical trials are shown with a marker indicating their publication type. This also aligns with a study from 2013 that logged and analysed data queries in a hospital and found that "[a]lmost a third of the articles [...] accessed were reviews." (Maggio et al., 2013). For quick orientation and relevance assessment, terms that appear in the patient's EHR are highlighted in the literature service. To help the literature relevance assessment process, a teaser text is displayed when hovering the mouse pointer over the eye icon after each publication title. In order to give the users a way to give feedback on the relevance of a publication and improve the literature search, icons with a thumbs-up and a thumbs-down are provided.

3.1.2 Evidence-Based Medical Recommendations

Evidence-based medicine describes the assessment and use of the best available research for decision making in patient treatment and diagnosis. This is done by focusing on well-designed, conducted research with a strong level of evidence like systematic reviews or randomized controlled trials (Hung et al., 2015). Existing web-based clinical decision support systems specialize on providing evidence-based treatment guidelines and summaries written by medical experts that reflect the current state of research. Obst et al. (2013) assessed the use of such a service named UpToDate.com and suggested that physicians often have little time and that navigating the service could at times be time consuming. They also noted that the summaries were sometimes written confusingly and that it took too long to quickly grasp the desired information.

The EBM recommendation service (Figure 2) therefore queries different secondary information sources for patient-related evidence-based summaries and reviews and extracts the most important sections that describe the patient's issue. If the users wish to see the extracted sections in context, they can follow the links and visit the original text.

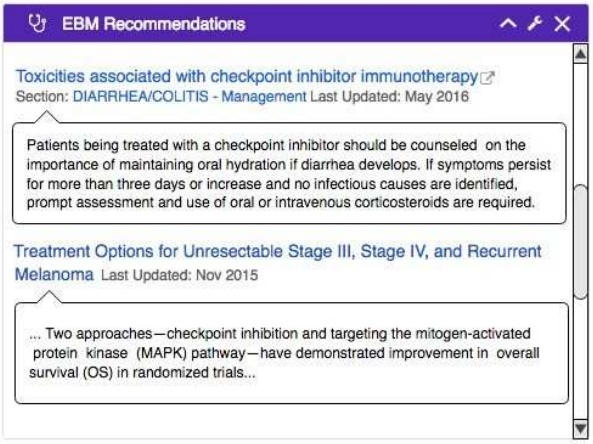


Figure 2. EBM recommendations service

3.1.3 Drug Information Service

Other important information in patient care is material on drugs and their interactions "at the point of drug prescribing" (Rahmner et al., 2012). Therefore, the drug information service provides information normally available in medication package leaflet inserts and secondary decision support services in a more accessible and structured way (Figure 3, left). The provided information includes dosage data for different age groups and pre-filled calculators to compute the correct dosage based on the age and weight of the patient. Other information consists of warnings, adverse effects, pregnancy, pharmacology, administration guidelines, material for patient education and pill images and prices. Selecting a drug for displaying can be done in an autosuggest-supported field that ranks already prescribed medication higher, but allows also searching for medication not yet prescribed.

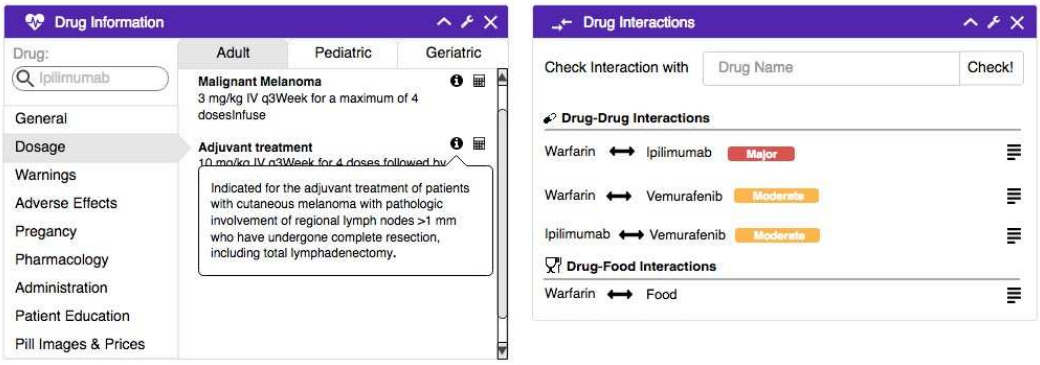


Figure 3. Drug information service

As physicians indicated they wanted to see automatically generated alerts for severe drug interactions and adverse effects (Rahmner et al., 2012), an alert is displayed prominently (Figure 3, top) . For more information on how to manage the interaction or alternative drugs, an appropriate link is provided. Non-severe drug interactions as well as drug-food interactions are displayed in an own panel where the users have the possibility to check interaction with other, not yet prescribed drugs (Figure 3, right).

To "make drug information more searchable" (Rahmner et al., 2012) and for example allow checking if a patient's symptom could be drug related, an adverse effects panel is introduced

(Figure 4). It automatically identifies drug-related comorbidities that are registred in the EHR but also allows searching for symptoms not yet recorded. An option to read more provides information on how to manage this effect, when it will occur or how long it will last.

Figure 4. Searchable Adverse Effects

3.1.4 Clinical Trials Service and News Service

Especially in cancer care the participation in clinical trials is an option for patients of all clinical stages as new findings lead to the development of many new drugs (National Cancer Institute, 2016). A clinical trials service is therefore introduced that searches for nearby clinical trials fitting the patient (Fiure. 5, left).

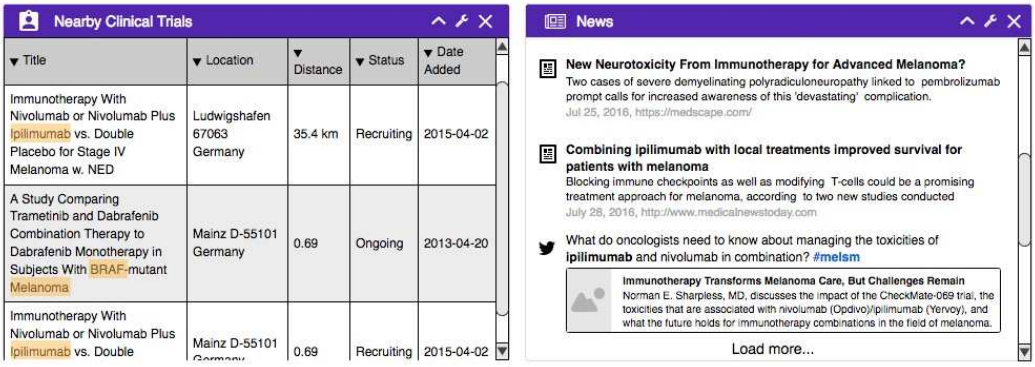


Figure 5. Clinical trials locator service and news service

Finally, a news service provides recent news on treatments, drugs, legislative information or other scientific breakthroughs that can be related to the current EHR (Figure 5, right).

3.2 Information Sources

In order to provide the CDSS services, several potential information sources were identified. The selection of information sources is a critical part in any CDSS application as they are the basis for the trust given them by physicians. As a patient's health might also depend on this information, the quality and correctness of the information is obligatory. This section introduces a summary of identified potentially relevant sources for CDSS.

3.2.1 Literature Service

There are an abundant amount of databases and search services for the health domain. Table 1 lists only a few of them that would be an option to use for the literature service.

Name	Description	API	Access	Size
Google Scholar	Search engine for scientific publications of all fields. Automatically crawls many journals.	no	commercial	estimated at 160 million articles
Ovid	Science search platform that includes many databases, including MEDLINE.	?	subscription	?
PubMed	Search engine mainly accessing MEDLINE database and focused on health topics. Query expansion by use of MeSH ontology.	yes	public & free	> 24.6 million records, about 500,000 new records each year
ScienceDirect	Website with access to large database of scientific publications from many fields.	yes	free (abstracts), subscription (full-text)	12 million records from 3,500 journals and 34,000 eBooks
Scopus	Database with abstracts and citations from many academic journals and many scientific fields, not focused on health topics.	yes	paid subscription	~55 million records
Springer API	Access to all Springer published journals, also includes BioMedCentral open-access publications.	yes	partly free, partly subscription	~2,000 journals and >6,500 books per year, access to >10 million online documents

Table 1. Literature data sources

3.2.2 Evidence-based Medical Recommendations

As evidence-based reviews or summaries should be written by medical experts to ensure quality and reliability, most services require a commercial licence or paid subscription to access them. Only a few public and free sources could be found in the scope of this thesis (Table 2).

Name	Description	API	Access	Volume
BMJ Best Practice	Evidence-based information to offer step-by-step guidance on diagnosis, prognosis, treatment and prevention.	yes	subscription	?
DynaMedPlus	Evidence-based clinical overviews and recommendations. Content updated daily. Also offers calculators, decision trees and unit and dose converters.	yes	subscription	> 3,200 topics and > 500 journals
EBMeDS	Platform-Independent web service CDSS with EBM module	yes	commercial	
Medscape / eMedicine	Largest clinical knowledge base available freely. Articles updated yearly. Also available as mobile application.	no	free, registration required	~6,800 articles
Physician Data Query	Cancer database from the U.S. *National Cancer Institute*. Contains peer-reviewed information on cancer treatment in the form of summaries for patients and professionals.	no	public	Only cancer domain
UpToDate	Popular evidence-based POC tool for a wide range of disciplines but targeted on internal medicine. Extensive peer-review process to ensure accurate and precise recommendations.	yes	subscription, some articles free	~8,500 topics

Table 2. EBM resources

3.2.3 Drug Information

There are few public resources for drug information that are also accessible via an API. Table 3 shows services that could be used in the context of the CDSS proposed in this work. Other not listed services include DrugBank (accessible over RxNav), ResearchAE.com, SIDER, NDF-RT, Epocrates or the OpenFDA API.

As Peters et al. (2015) stated, the data quality of some public resources might be a problem in clinical settings. They compared the two public drug interaction services DrugBank and NDF-RT and found a limited overlap between their drug interaction information. The commercial service used to compare against provided better coverage of the test data then both of the free services

combined. Assuming the commercial drug information services provide a better data quality, one would have to select one of those as data source. Additionally, as of September 2016, the public data source NDF-RT removed the drug interaction information from their service.

Name	Description	API	Access	Drug Information	Drug Interactions	Adverse Events	Drug Announcements/Recalls
DailyMed	Website by U.S. National Library of Medicine (NLM), provides high quality and up-to-date drug labels. Updated daily by FDA. Documents use structured XML format.	yes	public & free	✓	✓	✓	
MedlinePlus Connect	Service by NLM, provides unstructured natural language drug information/labelling and health topic overviews	yes**	public & free	✓*		✓*	
Medscape	Many clinical information resources available over website or mobile app. Articles updated yearly.	no	free, registration required	✓	✓	✓	
RxNav	Provides access to different drug resources like *RxNorm*, *NDF-RT* and *DrugBank*. Drug normalisation over different codes and systems by using *RxNorm*, drug interactions from *DrugBank*.	yes	public & free		✓		
Wolters Kluwer Clinical Drug Information	Commercial drug information APIs including interaction, adverse effects, indications and mapping to RxNorm.	yes	commercial	✓	✓	✓	

Table 3. Drug information sources

3.2.4 Clinical Trials

Several information sources for clinical trials were identified (Table 3).

Name	Description	API	Access	Type	Country
ClinicalTrials.gov	Trial registry from US National Institute of Health. 39% are U.S. only trials.	no*	Public & free	Register	Worldwide with a focus on U.S. (39%)
EU Clinical Trials Register	Clinical Trials Register for trials in the EU	no*	Public & free	Register	European Union
German Clinical Trials Register	German clinical trial register. Also imports trials from clincaltrials.gov that are located in Germany	no*	Public & free	Register	Germany
WHO International Clinical Trials Registry Platform	Search portal to central database with links to original records. Regular fetch of trials from currently 16 data providers, including sources mentioned earlier	no*	Public & free	Search Service	Worldwide

Table 4. Potential clinical trials sources

As the clinical trials' location plays a vital role in assessing their relevance, a service is needed that is not solely focused on one or a few countries. As the WHO registry platform seems to aggregate data from many national registers and other services, the choice would most likely fall on the WHO service.

3.2.5 News

Many online services offer medical news, e.g. MedScape, ScienceDaily or Medical News Today. Usually their content is not accessible over an API and there can also be legal issues prohibiting any “unauthorized copy, reproduction, distribution, publication, display, modification, or transmission of any part of this Service" (ScienceDaily, 2016).

An alternative to this might be the online social networking service Twitter which is used by many medical news sites and professionals.

3.3 Software Architecture

The CDSS is intended to be integrated into an EHR application. The application is organized in a three layer architecture where each layer consists of components that encapsulate logically separable units (Figure 5).

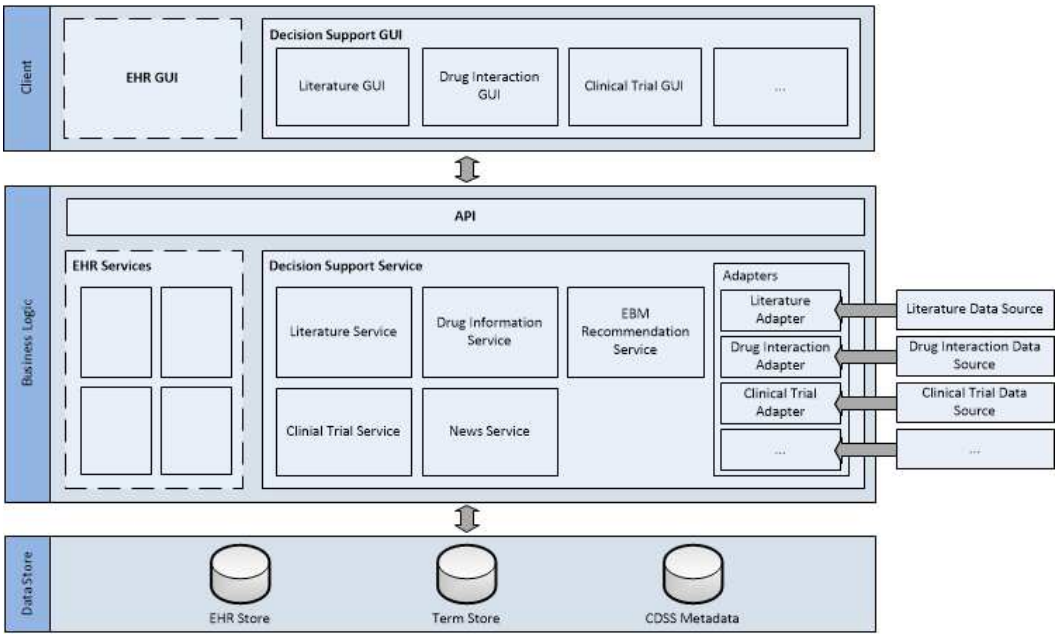


Figure 5. System architecture

On the client side, the Decision Support GUI is implemented alongside the EHR GUI. The different CDSS service are subcomponents of a single decision support module.

The business logic of the Decision Support System consists of the different service components as well as different adapters to connect to the information sources. The services store and use data from the EHR and the Term Store and persist data in the CDSS Metadata store.

3.4 Prototype Implementation

The literature service and drug information service have been implemented prototypically as part of an EHR application for melanoma treatment (Humm and Walsh, 2015) (Beez et al., 2015). The application is implemented in C# using .NET and MS SQL Server on the server side, and in HTML5 / CSS / JavaScript on the client side, using Bootstrap and AngularJS. As data source for the literature service PubMed was selected. Drug interaction data is acquired by querying DrugBank

over the RxNav drug interaction API. In the following sections, we describe the implementation of the literature service in more detail. See Figure 6 for the detailed architecture.

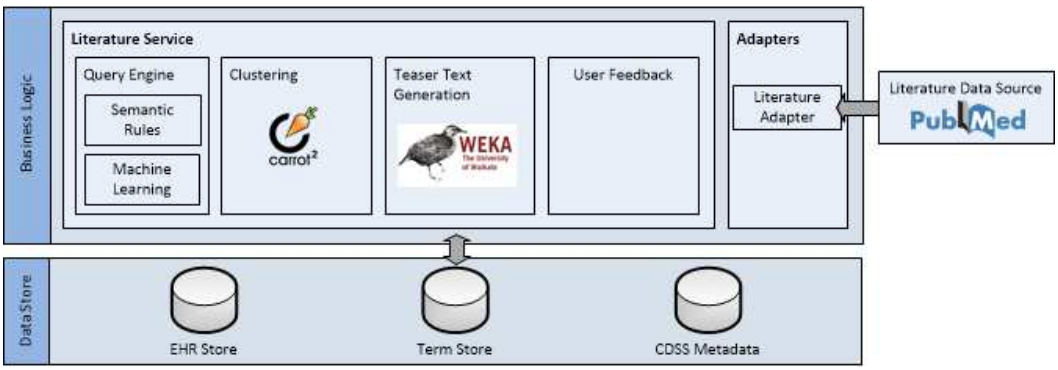


Figure 6. Literature service architecture

3.4.1 Query Engine

As PubMed is selected as the source for the literature service, we need a query in the literature source's query language to find and display publications. To generate the query from the EHR, two strategies are employed: use of semantic rules for creating queries from EHR attributes and machine learning.

Semantic rules

From the ca. 100 attributes that are currently used in the EHR application, not all are helpful for getting personalized literature suggestions for a concrete patient (Figure 7).

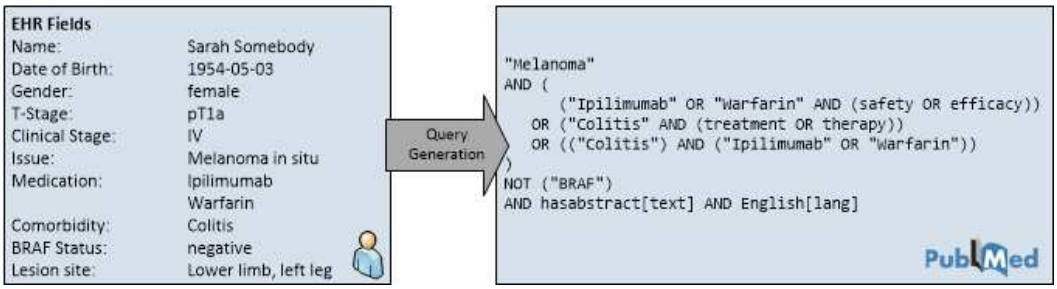


Figure 7. Sample query generation from an EHR

Fields with no relevance like the patient's name are omitted in query generation whereas relevant fields like the issue, medications or comorbidities are included using rules. One rule to search for publications that address the safety or efficacy aspects of one of the medications prescribed, combines all medication with an "OR" and adds "(safety OR efficacy)" to the medication subquery. Another rule combines the comorbidities field with the medication to search for drug-related adverse effects and their treatment. To ensure data quality and only search for recent literature, restrictions are added to the query like the "hasabstract[text]" to only show publications that contain an abstract.

Machine Learning Approach

The presented rule-based approach returns good results in various cases as fields like comorbidity, medication and issue are always relevant. However, in using a rules approach only a predefined field of possible questions can be answered and other fields like age, gender or the clinical stage could also in certain contexts return relevant literature. Additionally, the above mentioned terms "safety" and "efficacy" are only one example of additional query terms. Other terms that are not captured in the semantic rules could also become relevant. We therefore apply a query expansion method by automatically refining queries based on user feedback. This is done using a machine learning approach. Examples for additional query terms from the current prototype implementation include "adverse effects" and "drug therapy". Also, other attributes may become relevant in the future, e.g., the publication type "Review".

To facilitate machine learning, Accord.NET's [1] one-against-all multi label Support Vector Machine (SVM) is used to predict suitable search terms. Training data consists of user feedback data on the relevance of a literature given a specific EHR. As input vector, a bag-of-words (BoW) is built from all EHR terms. The output labels are extracted from the literature using the EHR's medical ontology.

Both strategies are used in parallel and their results are combined using an heuristic approach. The final query is passed to the literature adapter that retrieves the publications from PubMed and parses the retrieved XML document to an internal object representation. Publications are then ranked utilizing their PubMed rank in combination with explicit and implicit user feedback.

3.4.2 Clustering

Different users are interested in different topics and want to have different questions answered. A one-fit-for-all ranking is, therefore, not sufficient. To accommodate this, the result set is clustered to allow filtering the publications according to different semantic criteria. Example filter labels include "Ipilimumab-induced Colitis", "Adverse Effects" and "Overall Survival" (Figure 8).

Figure 8. Partial screenshot of automatically generated filter labels

For clustering the publications, the open source search engine clustering server "Carrot2" is used. It utilizes the specialized "Lingo" algorithm that automatically creates meaningful cluster labels from the publications' titles and abstracts (Osiński et al., 2004). The generated cluster labels are then filtered using the EHR application's medical ontology as well as a custom whitelist.

[1] http://accord-framework.net

3.4.3 Teaser Text Generation

If a publication title catches a consultant's attention, he may choose to read a teaser text in order to easily assess its relevance. This teaser text should be about three sentences long and contain the conclusion of the publication. Usually, abstracts consist of at least 500 or more words and not all of them have their conclusion displayed in a structured form. Therefore, for publications for which the conclusion is not explicitly marked, a machine learning algorithm is employed to predict the concluding sentences. The implementation uses the libSVM algorithm and the open source machine learning software Weka.

3.4.4 User Feedback

There are two kinds of user feedback gathered in this application, active and passive. Active feedback is generated by the users clicking on thumbs-up or thumbs-down icons in the client. Passive feedback is gathered by logging and interpreting all clicks on a publication. All feedback is stored in the CDSS Metadata store and is used for the ranking of publications and clusters as well as for creating training data for machine learning.

4 Evaluation

We compare concept and the prototypical implementation with the requirements stated in Section 2. Requirement 1.2 (personalized) and 1.3 (pro-active) are obviously met since the CDSS information is displayed automatically based on the EHR currently being handled. Also, Requirement 1.5 (Workflow) is met, since the CDSS panels can be separated from the EHR dialogs. Assessing the Requirement 1.1 (relevant), 1.4 (easily comprehensible) and 2.1 (usable) is less obvious and needs feedback from users. Therefore, we have conducted an initial survey with 4 medical students and 1 resident physicians.

Users were given the task to prepare for a multidisciplinary team (MDT) meeting using the different CDSS modules. Subsequently, they were to fill out a questionnaire to assess the usability aspects and relevance of the CDSS services implemented in the prototype as well as in the interaction concept.

Initial usability observations indicate positive feedback. Positively mentioned was the display of the teaser text for assessing literature relevance, the drug interaction information, searchable adverse effects as well as the automatically generated filters. The usability test also revealed some weak points that were, subsequently, improved. For example, the teaser text's icon position was initially not prominent enough and was easily overlooked. Therefore, its position was moved after the literature's title and a legend was added (Figure 1). However, the survey was only an initial one and relevance assessment by medical students might be skewed due to their lack of clinical experience. In the future, a comprehensive survey with physicians working in the melanoma domain is planned.

When opening EHR for the first time initial loading of literature data may take up to 15 seconds. However, as this happens asynchronously while the consultant is working with EHR this will not interfere with the EHR workflow. As soon as all data is loaded and the physician enters the CDSS, each interaction is less than 500 ms which meets Requirement 2.2 (low response time) clearly. With caching strategies this initial loading time may also be reduced.

Concerning Requirement 2.3 (extensible), the component-based system architecture and the use of adapters to access information sources enables the extension of the CDSS and implement other decision support modules with moderate implementation effort. For a new data source one would have to implement an additional adapter which results in about 100 lines of code (LOC). For a new decision support module the service on the server and the client GUI would have to be implemented. For the client this would require about 300 LOC. The server implementation depends on the module's complexity, e.g. the literature service is ~2,500 LOC whereas the interaction service is ~200 LOC.

5 Related Work

The idea of CDSS is not new and there exist many services accessible over the browser and/or smartphone applications. Their scope ranges from drug information, drug interaction and diseases to guidelines, pill identification and alternative medications. Example services include UpToDate, Epocrates, MedScape and First Databank. Often, the integration of theses services into an EHR system consists of providing a standard search field that enables the users to search and visit the CDSS service's main webpage. Some EHR systems like Athenahealth's EHR include context-sensitive drug monographs that provide information like dosing, adverse effects and other key safety data directly in the EHR [2]. However, there are systems that include include the patient's context in the CDSS search. The integration of UpToDate into various EHR systems provides such a future by displaying an info button in various EHR locations and providing an automatic search [3]. However, this function delivers standard UpToDate results pages and problems mentioned earlier in this work like confusingly written texts and difficulties navigating the long summaries remain (Obst et al., 2013).

Finland's Evidence based Medicine electronic Decision Support (Nyberg, 2012) is a platform-independent online service that accepts structured EHR data as input and returns clinical decision support data like links to guidelines, therapeutic suggestions, clinical reminders and alerts. These rules can be created by experts in a web-based editor and scoped per organization or globally. It can also populate forms and calculators with patient specific data. They do not include a literature search service but provide other services like the drug alerts that are similar to services presented in this work. Additional services like the knowledge assistant would be relevant in the context of this work and could later be integrated into our proposed CDSS.

Alternative approaches for the task of finding literature fitting to a patient's case has been described in different publications. Perez-Rey et al. (2012) propose a visual browser extension that allows the user to select a subset of extracted search terms from a natural language medical record. These selected terms will then be used to search PubMed. However, the selection of search terms is not automatic or pro-active as the user has to interact with the application to build the search.

[2] http://www.athenahealth.com/enterprise/epocrates/clinical-decision-support

[3] http://www.uptodate.com/home/hl7

Soldaini et al. (2015) propose a CDSS that tries to find fitting medical literature for medical case reports instead of EHRs. They consider the natural language case report as the query and apply query reformulation techniques like query reduction by identifying medical terms and expansion by using pseudo relevance feedback to build the search. They try to best answer the case report (~60 words) by providing relevant literature. In contrast to this work they do not use PubMed as search engine but use a local search server.

To the best of our knowledge, no system integrates a literature search system into an EHR to display relevant medical literature. Similarly, we believe there is no implementation of a patient-specific search service for clinical trials.

6 Conclusions and Future Work

Personalised medicine offers great promises for consultants' decision making, potentially resulting in improved patient treatment. Numerous medical information sources are already available which can be utilized, and they are growing rapidly. However, personalized medicine has not yet widely used in day-to-day clinical practise. With the concept and a prototypical implementation of a Clinical Decision Support System (CDSS) for personalized medicine presented in this paper, we intend to make a contribution towards this direction.

Particular focus has been on the usability of the CDSS allowing consultants to quickly and intuitively gather relevant information with minimal system interaction. A number of artificial intelligence (AI) techniques have been used to tailor information to the patient's medical situation and the consultant's information demand.

It is planned to integrate the CDSS presented into a commercial EHR application suite for melanoma treatment. Towards this end, future work is required. Additional information services as presented in the interaction concept need to be implemented. A comprehensive analysis of the CDSS with consultants in the field needs to take place resulting in potential improvements of the concept and the implementation. Then, a trial phase with real patient data is planned to extend the data base for machine learning. We intend to publish insights from these analyses.

May this work eventually help consultants improve patient care.

7 Acknowledgements

This work was funded by the European Commission, Horizon 2020 Marie Skłodowska-Curie Research and Innovation Staff Exchange, under grant no 644186 as part of the project SAGE-CARE (SemAntically integrating Genomics with Electronic health records for Cancer CARE).

8 References

Academy of Medical Sciences (2015), Stratified, personalised or P4 medicine: a new direction for placing the patient at the centre of healthcare and health education (Technical report). Academy of Medical Sciences. May 2015. Retrieved 24/8/2016.

Beez, U., Humm, B.G. and Walsh, P. (2015), "Semantic AutoSuggest for Electronic Health Records", 2015 International Conference on Computational Science and Computational Intelligence (CSCI), 760-765.

Humm, B.G. and Walsh, P. (2015), "Flexible yet Efficient Management of Electronic Health Records". 2015 International Conference on Computational Science and Computational Intelligence (CSCI), 771-775.

Hung, B. T., Long, N. P., Hung, L. P., Luan, N. T., Anh, N. H., Nghi, T. D., … Hirayama, K. (2015), “Research Trends in Evidence-Based Medicine: A Joinpoint Regression Analysis of More than 50 Years of Publication Data”, PLoS ONE, 10(4).

Kawamoto, K, Houlihan, C.A., Balas, E.A. and Lobach, D.F. (2005), "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success", BMJ? British Medical Journal, 330(7494), 765.

Maggio, L. A., Cate, O. T., Moorhead, L. L., Van Stiphout, F., Kramer, B. M., Ter Braak, E., Posley, K., Irby, D. and O'Brien, B. C. (2014), “Characterizing physicians' information needs at the point of care”, Perspectives on Medical Education (Perspect Med Educ), 33(5), 332-42.

Maggio, L. A., Steinberg, R. M., Moorhead, L., O'brien, B. & Willinsky, J. (2013). "Access of primary and secondary literature by health personnel in an academic health center: implications for open access". Journal of the Medical Library Association (J Med Libr Assoc), 101(3), 205-12.

Marchant, G. E. and Lindor, R. A. (2013), “Personalized medicine and genetic malpractice”, Genetics in Medicine (Genet Med), 15(12), 921-2.

National Cancer Institute (2016). Melanoma Treatment. URL: http://www.cancer.gov/types/skin/hp/melanoma-treatment-pdq#section/_885 (visited on 11/09/2016).

Nyberg, P. (2012), “EBMeDS Clinical Decision Support. EBMeDS White Paper”, URL: http://www.ebmeds.org/www/EBMeDS%20White%20Paper.pdf (visited on 12/09/2016).

Obst, O., Hofmann, C., Knüttel, H. and Zöller, P. (2013), “‘Ask a question, get an answer, continue your work!’ – Survey on the use of UpToDate at the universities of Freiburg, Leipzig, Münster and Regensburg”, GMS Medizin—Bibliothek—Information, 13(3), 26.

Osiński, S., Stefanowski, J. and Weiss, D. (2004), “Lingo: Search results clustering algorithm based on singular value decomposition”. In Intelligent information processing and web mining, 359-368, Springer Berlin Heidelberg.

Perez-Rey, D., Jimenez-Castellanos, A., Garcia-Remesal, M., Crespo, J., & Maojo, V. (2012), “CDAPubMed: a browser extension to retrieve EHR-based biomedical literature”, BMC Medical Informatics and Decision Making, 12, 29.

Peters, L. B., Bahr, N. and Bodenreider, O. (2015), “Evaluating drug-drug interaction information in NDF-RT and DrugBank”, Journal of Biomedical Semantics, 6, 19.

Rahmner, P. B., Eiermann, B., Korkmaz, S., Gustafsson, L. L., Gruvén, M., Maxwell, S., … Vég, A. (2012), “Physicians’ reported needs of drug information at point of care in Sweden”, British Journal of Clinical Pharmacology, 73(1), 115–125.

ScienceDaily (2016). Terms and Conditions of Use. URL: https://www.sciencedaily.com/terms.htm (visited on 11/09/2016).

Soldaini, L., Cohan, A., Yates, A., Goharian, N. and Frieder, O. (2015), “Retrieving medical literature for clinical decision support”, European Conference on Information Retrieval, 538-549, Springer International Publishing.

Prototype proof of concept for a mobile agitation tracking system for use in elderly and dementia care use cases

Michael Healy, Alfie Keary, Paul Walsh

Department of Computer Science

CIT – Cork Institute of Technology, Ireland

michael.healy2@mycit.ie; alphonsus.keary@cit.ie; paul.walsh@cit.ie

Keywords

Vision, Agitation, Affective, Dementia, CMAI

Abstract

The MAT project aims to develop and evaluate an initial set of algorithms that can detect agitation, restlessness and aggression in dementia patients. MAT uses vision based analytics to track a subjects facial expressions in real-time. The first version of MAT has implemented two use cases for the detection of restlessness and aggression. The project also has the potential to be used in more advanced machine learning and data analytics applications typically for research purposes on elder care. Data sets for each subject's monitoring period are generated in CSV file format which could be used to populate a database or as inputs to machine learning classification algorithms/platforms.

1 Introduction

Dementia is a chronic or persistent disorder of the mental processes caused by brain disease or injury and marked by memory disorders, personality changes, and impaired reasoning. It describes a wide range of disorders including Alzheimer's disease, which is the most common type of dementia and makes up 50% to 70% of dementia cases (WHO, 2014). Other common forms of dementia include vascular dementia, Lewy body dementia and frontotemporal dementia. Globally, dementia affects 36 million people (WHO, 2014) and it is estimated that 47,983 people were diagnosed with dementia in Ireland in 2011 (Pierce, Cahill and O'Shea, 2014). There is no known cure for dementia. This means that dementia patients require full time caring which can put a lot of pressure on their care givers. Research has

A study conducted on 408 nursing home residents found that there is a direct link between agitated behaviours and cognitive function (Cohen-Mansfield, Marx and Rosenthal, 1990). The study used the Cohen-Mansfield Agitation Inventory (CMAI) (Cohen-Mansfield, 1991) which is used to score 29 different agitation states on a 7-point scale. These 29 states fall under three categories;

- **Aggressive behaviour,**
- **Physically nonaggressive behaviour,**
- **Verbally agitated behaviour.**

The CMAI describes some of the physically non-aggressive behaviours such as pacing, stealing, trying to get to a different place, handling items inappropriately and general restlessness. The report also found that "*cognitively impaired patients were significantly more physically non-aggressive than patients with low cognitive impairment*" (Cohen-Mansfield, Marx and Rosenthal, 1990).

CMAI is a complex evaluation process to be carried out by trained professionals and it has been used as the foundation for the MAT project from a clinical perspective. The examples listed below

(*relating to typical CMAI based visual observations of a subject*) have been the basis for the initial set of vision based algorithms to be developed for MAT.

- Agitation may be abusive, aggressive or repetitive behaviour towards self or others such as shouting, head banging, staring and repetitive mouth opening and closing.
- Agitation may be appropriate behaviour performed with inappropriate frequency, such as excessive head movement, rocking, trying to get up from a chair or out of bed constantly, and asking the same questions continually.
- Agitation may be inappropriate according to social standards for the specific situation, as in spitting at a person, hitting a person or taking clothes off in a common room.

The Mobile Agitation Tracker (MAT) project aims to develop and evaluate an initial set of algorithms that can detect restlessness and agitation generated aggression in elderly patients and dementia care subjects. Cognitively impaired elderly subjects often show signs of being in an agitated state. The Merriam Webster medical dictionary defines agitation as a "state of excessive psychomotor activity accompanied by increased tension and irritability" (Merriam-Webster.com, 2011). People who suffer from agitation are often tense, confused, and irritable. They may also show signs of aggressive behaviour. Research has already showed that affective computing could be applied to manage and care for the emotional wellbeing of elderly people (Bond et al., 2016). The first version of MAT has been developed to address the following two typical use cases.

Use Case #1: Restlessness Detection.

- MAT monitors the head movements of an in-camera subject for a given time period to detect if the subject may be in a restless state or not. This typically relates to a subject that may be in bed or confined to a chair that is demonstrating excessive head movement which may be any mixture of head rolling, head banging of an object, excessive tilting or drooping of the head. MAT also provides a report for a subject's carer for on-going analytics on head movement based possible agitated behaviours.

Use Case #2: Aggression Detection.

- The second MAT use case is more advanced and involves the monitoring and tracking of specific facial landmark points from the in-camera facial frame images of the subject. This use case is closely related to the CMAI visual observation use case one outlined above and is designed to detect if a subject is showing signs of a violent outburst or abusive behaviour. This version of MAT is able to detect facial expression images that demonstrate that the subject is showing an aggressive or angry facial expression.

1.1 Review of State of the Art

Intel Real Sense

To develop an algorithm for visual processing, a camera that is capable of tracking movements and positions of the human face is necessary. Intel's Real Sense camera was one solution that was considered for use with this project. The camera is made up of three cameras that act as one, a full 1080P x 1920 camera, a VGA depth resolution camera and an IR projector. IR projectors use infra-red light points that are projected onto the users face or body. The VGA camera can than detect these points and use the differentiation of these dots to determine depth and distance between them. The Real Sense camera has a range of 0.2 – 1.2 meters and it also has dual microphones for

capturing audio. A single camera is capable of tracking up to five people at one time. This could be very usefully in a care home scenario when multiple patients are in shared rooms or locations. The camera comes with an SDK for developing applications on the platform. The foundation of the SDK stack is made up of the SDK core, I/O module and the algorithm modules. The SDK core manages the application pipeline execution and the I/O modules of the camera. The algorithm modules acts as middleware for the main functions of the camera such as hand tracking, face tracking and gesture recognition. They also allow development of applications through development frameworks (C++, C#, Java, Unity). A diagram of the SDK can be seen in Figure 1.

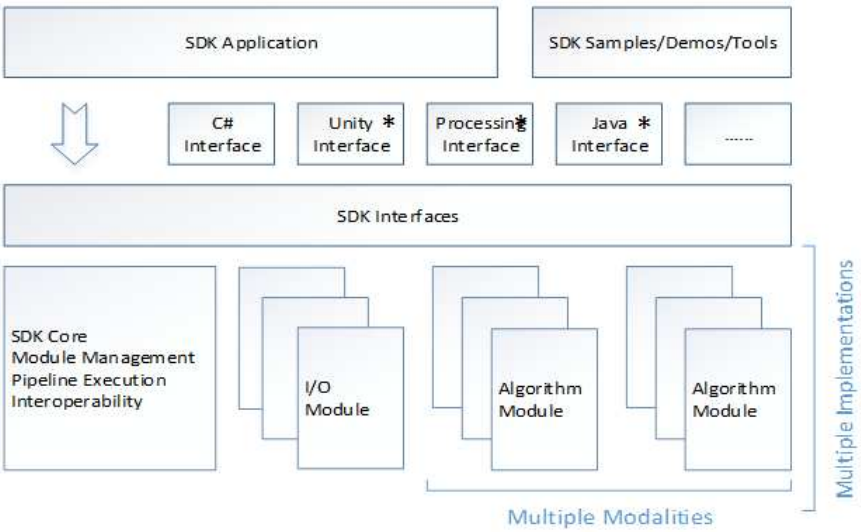


Figure 1. Intel® Real Sense™ SDK Architecture

Microsoft Kinect

Another option that was considered was Microsoft's Kinect which is also a vision based gesture recognition camera that was originally developed for the Xbox 360 game console but there is now an SDK available for Windows operating systems. The Kinect is very similar to the Real Sense camera, it has a 1080P colour camera, a VGA depth camera with a range of .5 – 4.5 meters and an IR projector camera that works the same as the Real Sense camera. The Kinect is particularly good at body tracking. It can track 19 different points on the human body from the feet to arms and head. The Kinect also comes with an SDK to allow developers to develop third party applications to run on a windows operating system. This limits the choice of software languages to C++, C# .NET or Javascript with HTML/CSS. There are some other limitations to the Kinect, it only allows for two people to be tracked at one time, the recommended tracking distance from person to camera is 6 feet. This can become a problem when patients may be in small rooms with limited space. A diagram of the Kinect SDK can be seen in Figure 2.

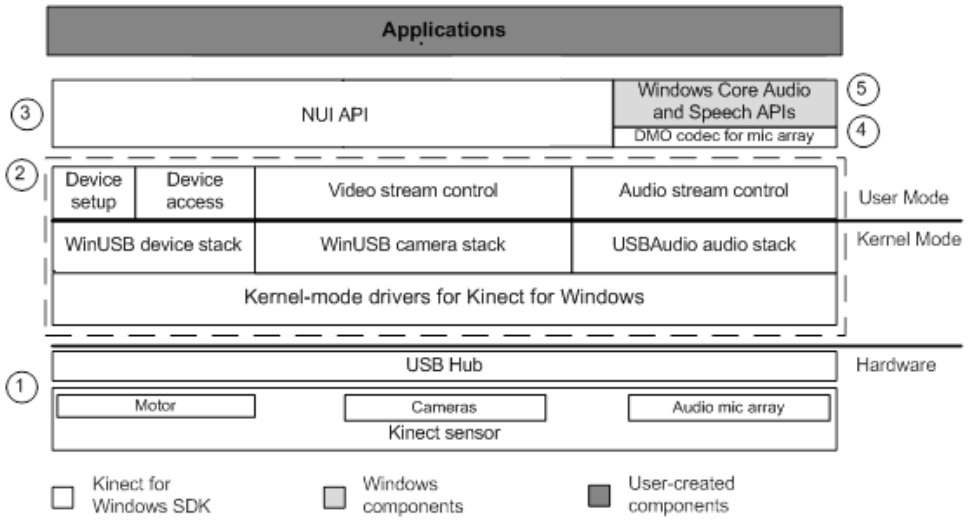


Figure 2. Microsoft Kinect SDK Architecture

The Chosen Solution

For this project the decision was made to use the Intel Real Sense camera. This decision was based on a number of factors. The Kinect camera was mainly developed for body tracking and doesn't feature the same suite of algorithms for analysis of head, hands or face that the Real Sense SDK provides. This also means that the Kinect was designed for the user to be a large distance away from the camera which could be an added constraint when possible locations for the implementation of this project were taken into account. The Real Sense camera was designed to be used closely to the patient which allows for more detailed analysis. For example, the camera has face analysis modules which provides the ability to recognise and track 77 different points on the human face. This would be particularly useful for developers when attempting to determine if the patient is showing signs of aggression. The final consideration to take into account was the form factor of both cameras. The Kinect is a large camera that requires either a table or a stand. The Real Sense camera is a much smaller size and can be easily fitted on to any monitor or laptop screen. The Real Sense camera is also available built into some new laptops that are produced by HP, Acer and Dell (Dell, 2016).

2 Solution (Material & Methods)

The main functionality of the project was implemented as a Microsoft Windows™ application. This application interacts with the Real Sense camera to monitor patients, analyse their behaviour and movements and decide whether to alert a care worker or not. The application was written in C++ and developed using Visual Studio 2013. The model of Real Sense camera used in the implementation is the F200 and the SDK version is R4 (2015).

The MAT application doesn't have a GUI, as it is a console only application. This was because of the time constraints of the project, however future developments of the application could allow for the implementation of a graphical user interface. When the application loads the user is prompted to enter a defined time period in seconds. This time period represents how long the system should capture image frames for before the visual analytical process begins. The system contains two objects, one "restlessness" object and one "aggression" object. These objects are used to store information from the visual frames so that the M.A.T system will be able to determine if the patient is restless or aggressive. The Real Sense camera has the potential to capture up to 60 frames per

second. For the scope of this project however, the maximum frames we will capture is 30 per second. This means that the system is capturing data from unique frames up to 30 times a second. The next section will look at the three types of data structures that are captured and stored from each frame.

2.1 Expression Data

Inside the restlessness object there are vectors of "ExpressionsData". Expression data is what gives us information from the camera about the facial expressions of the user. It has two main components, the actual "ExpressionsData" interface and "FaceExpressionResult". The SDK allows us to query a number of predefined expressions. After an expression query has been made, you can also query some attributes from the results. The attributes vary depending on what is queried but the main one is intensity. The intensity can be used for example, to see how wide a mouth is open or how closed an eye is. For example we could use this information to determine if a user is sleeping if both eye closed intensities are 100%. The MAT system captures expressions of eyes, mouth and smile and stores this information in the restlessness object.

2.2 Pose Data

The definition of a pose is to "*assume a particular position in order to be photographed, painted or drawn*" (dictionary.com, 2016). M.A.T carries out pose analytics using foundational built in Real Sense camera and SDK supports. The camera can retrieve head position information such as yaw, pitch, and roll of the users head. Yaw, pitch and roll are measurements of rotation of an object in a 3D space. It can be clearly visualized in Figure 3 below. Each restlessness object also contains a vector of pose data for each frame.

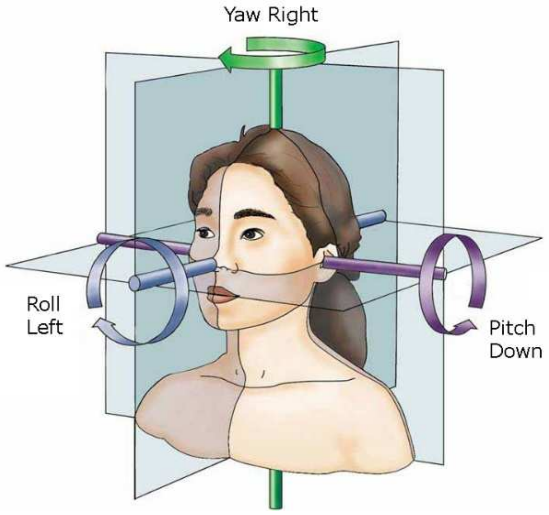


Figure 3. Pitch, Yaw & Roll of Human Head (IMG, 2016)

2.3 Landmark Data

Facial landmarks are defined as the detection and localisation of certain key points on a human face. They are also known as vertices or anchor points. Current RealSense cameras can accurately distinguish 77 landmark points on a human face. The points are grouped around the following areas: the eyebrows, eyes, nose, mouth and finally all the way across the jaw line. Figure 4 below shows all the landmark points on a face. The aggression object contains vectors of these landmark points. It stores data for point "0" (left tip of the right eyebrow), "5" (right tip of the left eyebrow) and "29"

(tip of the nose) for each frame captured. Intel's Real Sense camera allows the location of these points in a 2D or 3D space to be retrieved (x, y and z coordinates).

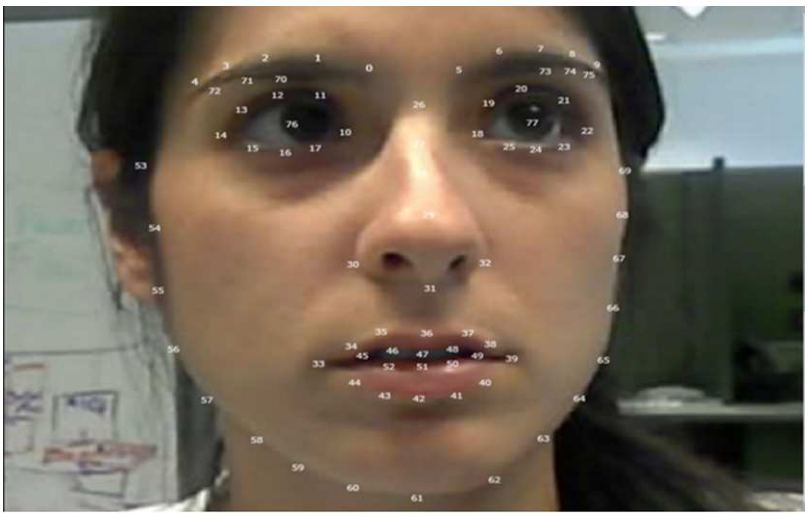


Figure 4. Intel® Real Sense™ SDK Landmark Data

Once the "defined time" period entered by the user has passed, the restlessness object and the aggression object is passed to the "analyser" class which then tries to determine if the data in the objects passed show signs that the user may be either restless or aggressive. The analyser is run on a separate thread so that the capturing of frames isn't interrupted. The next section of this chapter explores how the analyser works and shows the code behind it.

2.4 Analysing Data

The analyser is where the system implements the main functionality. The analyser class contains 5 variables that are passed from the main class. These variables are a restlessness object, an aggression object, the ID and name of the patient that was monitored and finally the amount of time in which these frames where captured over (this was defined by the user at the start of the application). The analyser also contains some predefined variables. These variables are used in the analysis of the frames and can be seen in Figure 5. For example the "YAW_RADIUS_POS" defines that the users head must be turned at least 13 degrees to the right before they are considered to be looking right by the system. The analyser then proceeds to check for restlessness or aggression.

```
#define YAW_RADIUS_POS 13                // Point (in degrees) at which head must be turned before it is deemed to be looking right.
#define YAW_RADIUS_NEG -13               // Point (in degrees) at which head must be turned before it is deemed to be looking left.
#define DIFFERENCE_NOT_MORE_THEN 30      // Difference in percentage of time looking left/right should not be more then defined.
#define DIFFERENCE_NOT_LESS_THEN 1       // Difference in percentage of time looking left/right should not be less then defined.
#define PERCENTAGE_MOVING 40             // Percentage of time spent not looking straight.
#define ACCURACY 100;                    // Multiplication value to increase value of distance coordinates
#define EYEBROWTONOSE 4.8                // Decrease of distance from eyebrows to nose;
#define EYEBROWTOEYEBROW 2.1             // Decrease of distance between eyebrows;
#define PERCENTOFFRAMES 10               // Percentage of aggression frames needed to flag alert.
```

Figure 5. Predefined Variables

2.5 Is the patient Restless?

The first part of this method is to count how many frames the patient was looking left or right. This is achieved using the pose data. It queries the yaw value for each frame and checks that the value for yaw is within the predefined values for looking left or right. After this has been calculated, the

program calculates what percentage of frames the user was deemed to be looking left or right. Percentages are calculated because we don't know how many frames will be captured within the given time. The frame rate of the camera can vary from 20 frames per second up to 30 frames per second. This prevents calculations from being skewed at a later stage in the method.

Next the system checks if the percentage of time looking right added to the percentage of time looking left is greater than the value defined in "PERCENTAGE_MOVING". This defines that the head must be moving for e.g. 50% of the time the patient was monitored for. If this condition is not true the system concludes that the patient was not restless and returns false. Otherwise the system proceeds to perform more analysis on the frames. The system now looks at the percentage of time the patient was looking right. If it is greater than the percentage of time looking left and the difference between the time looking left and right is not greater or smaller then values that were predefined then the users head is said to have been shaking. The same condition is checked in reverse if the percentage of time looking left was greater than looking right. The same check is done on the pitch of head to see if the patient was rocking their head or attempting to move. The parameters "not greater than" or "not less than" are used to distinguish the patient simply looking to the left or right in a normal fashion or just a light nod of the head from actual head shaking. If these conditions are true than a "true" Boolean value is returned, otherwise a "false" value is returned. In future implementations, machine learning techniques such as neural net or SVM could greatly improve the accuracy of detection.

```
if ((percentageRight + percentageLeft) > PERCENTAGE_MOVING)
{
    if ((percentageRight >= percentageLeft) && (percentageRight - percentageLeft<DIFFERENCE_NOT_MORE_THEN)
        && (percentageRight - percentageLeft>DIFFERENCE_NOT_LESS_THEN))
        return true;
    else if ((percentageLeft > percentageRight) && (percentageLeft - percentageRight<DIFFERENCE_NOT_MORE_THEN)
        && (percentageLeft - percentageRight>DIFFERENCE_NOT_LESS_THEN))
        return true;
    else
        return false;
}
else
    return false;
```

Figure 6. Calculating Head Shaking

2.6 Is the patient aggressive?

The next check the system performs is looking for aggression in patients. It does this by looking at landmark points. The first step is to initialize a counter that will add how many frames the system classifies as showing signs of aggression. The next step is to measure the distance from the tip of the left eyebrow to the tip of the nose. To do this we need to get the (x, y) coordinates and ensure these coordinates are accurate. This can be done using the "confidence" variable that is in each landmark object. The confidence variable outlines how confident the camera feels that the values are accurate. If the confidence value is greater or equal to 90% the system accepts the data. After the system validates that it is confident the values are accurate, it assigns the x and y variables and passes them into a method called "findDistance". This method uses geometry to find the straight line distance between the two points. The formula used to achieve this is the straight line distance formula which returns the Euclidean distance between two coordinates. The full formula can be seen in Figure 7.

$$\sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$

Figure 7. Euclidean Distance

This process is then repeated to get the distance from the right eyebrow to the tip of the nose and final the distance between the tips of both eyebrows. If all the distances are reduced to the level that is predefined in the program then the aggressive frame counter is increased by one. Finally, after all the frames have been analysed, the find percentage method is used again to determine out of all the frames, how many of them did the patient show signs of aggression. If the percentage is greater than the amount predefined than the patient is said to be showing signs of aggression and a Boolean value of "true" is returned. Otherwise a value of "false" is returned

2.7 Generating CSV file

A Comma Separated Value File (CSV) is a file type that allows data to be saved in a table structured format. These files can be easily read with software such as Microsoft Excel, or used to populate databases. After the frames are finished being analysed the program outputs all the captured frame data to a CSV file. This file contains values for all the expressions and landmarks. It is hoped that this information could later be used in machine learning algorithms to develop models that could outline different emotional states such as aggression, or used to better define behaviours such as restlessness. They could also be used by experts to further analyse the data to spot trends in the patient's behaviours. The program passes both the restlessness object and the aggression object into the method that generates the CSV file. The current date and time is captured and is used for the filename. This ensures that all filenames are unique and no files will overwrite each other. The program then outputs the patients ID, all the intensity values for each expression and the X, Y coordinates for the landmark points. Each row of the file accounts for a frame captured. The longer the monitoring period, the more rows will be in this file. It will also contain much more information.

	A	B	C	D	E	F
1	Patient ID	11				
2	Time Generated	02/05/2016	14-50-10			
3	Right Eye	Left Eye	Mouth	Pitch	Yaw	Roll
4	0	0	0	3.66359	-6.59658	-1.22292
5	0	0	0	3.17327	-5.3304	-0.150679
6	0	0	0	2.63093	-5.44341	0.114227
7	0	0	0	2.76178	-4.75474	0.145633
8	0	0	0	2.17967	-4.92214	0.00619111
9	0	0	0	1.55811	-5.05076	0.0204711
10	0	0	0	1.05839	-5.21329	0.125868

Figure 8. Output of CSV File

3 Evaluation

MAT has originated from a final year computer science project and has been supervised by researchers from the Cork Institute of Technology, SIGMA group (Cork Institute of Technology (CIT), 2016). The system has been tested in a lab based environment for both the vision and cloud based alert aspects of the software. Both use cases have been tested extensively within the lab but due to ethical, security and privacy considerations the MAT platform has not reached a stage where

it is ready for real-world testing. Also during the testing phase the following limitations were identified in relation to the current versions of the detection algorithms. For example, performing an analysis of a subject's mouth to distinguish normal speaking from shouting or screaming is challenging and needs further work. Also this could be enhanced in a future version of MAT by incorporating voice analytics into the detection algorithm. The algorithm for finding common facial features that could be used to identify aggression in subjects is also at a very early stage and could be improved using affective computing techniques (Picard, 1997), facial expression databases and machine learning algorithms.

The following screen shot demonstrates the lab based testing of the real-time tracking capabilities of the MAT platform.

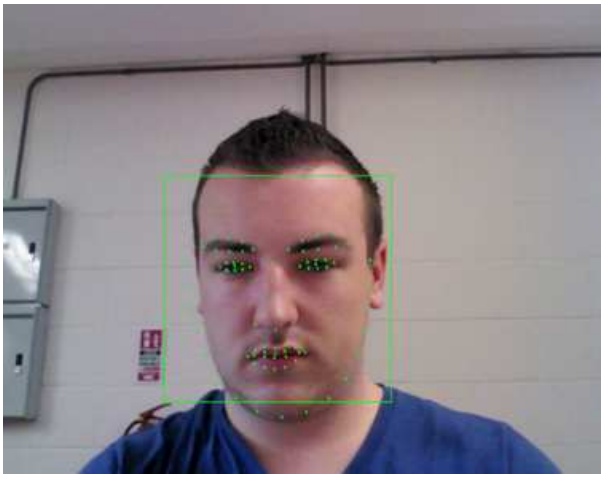


Figure 5. Example of facial landmarks generated by MAT

4 Conclusion and future work

The M.A.T research project has proved that the development of an algorithm, that could detect restlessness and aggression is possible by utilizing advanced vision based gesture recognition systems. This project has demonstrated this with the use of piloting and test cases. This project has proved that carer and also perhaps insurance costs could be significantly reduced and overall quality assurance of care could be increased by using the M.A.T technology to monitor patients in real-time either in their own homes or in care establishments.

The Cork Institute of Technology is the lead partner in the development of an affective computing based platform called SenseCare (SenseCare EU Partnership, 2016, www.sensecare.eu). SenseCare is a first of its kind and will uniquely create a cloud based affective computing platform capable of processing and fusing multiple sensory data streams to provide cognitive and emotional intelligence for AI connected healthcare systems (Keary, Alphonsus, and Walsh, 2014). The underlying sensory layer of the SenseCare platform involves the fusion of data streams from multiple affective sensory devices such as vision, wearables, audio and EEG devices. These data streams are then processed by a fusion and machine learning layer in the SenseCare platform that delivers cognitive and affective state data on a subject.

MAT has the potential to be incorporated into the lower layers of the SenseCare platform as it can provide raw facial landmark data that can be fused with the other affective based sensory adaptors configured for the SenseCare platform. Currently MAT provides data sets for each subject's

monitoring period in a CSV file format which could be used to populate SenseCare data repositories or act as a feature set for SenseCare related machine learning classification algorithms.

MAT is at an early stage of development but the next version of the software will see increased focus and implementation of other aspects of the CMAI such as analysing the subject's body movement as mentioned in the visual observations above. This would see the MAT project addressing further aspects of CMAI process and would greatly improve accuracy in detecting agitation. Also machine learning algorithms (possibly SenseCare based) could be used to analyse both existing data and data created by MAT. This analysis would provide advanced insights into agitation detection for elderly care and future research. Work is also expected to start on the development of a MAT based adaptor for the SenseCare platform. This work will lead to further development of the existing MAT algorithms with a specific SenseCare related focus.

5 Acknowledgment

I would like to sincerely thank the SIGMA research team for their supervision, patience, attention and guidance throughout the project. Without their expert knowledge this project would not have been possible. Alfie Keary is a funded researcher on the H2020 EU funded MSCA RISE project SenseCare, grant No. 690862

6 References

Bond et al. (2016). "SenseCare: Using Affective Computing to Manage and Care for the Emotional Wellbeing of Older People" in EAI International Conference on Wearables in Healthcare, Budapest, Hungary

Cohen-Mansfield (1991). "Cohen-Mansfield Agitation Inventory" American Psychological Association

Cohen-Mansfield, Marx and Rosenthal (1990). "*Dementia and Agitation in Nursing Home Residents*". s.l. : American Psychological Association

Dell (2016). Venue 8 7000 Series Android™ Tablet | Dell . [ONLINE] Available at: http://www.dell.com/en-us/shop/productdetails/dell-venue-8-7840-tablet. [Accessed 30 October 2016].

dictionary.com (2016). *Pose | Define Pose at Dictionary.com*. Available at: http://www.dictionary.com/browse/pose. (Accessed 29 October 2016).

IMG (2016). *Image adapted from a Leeds University medical examination* at mcqs.leedsmedics.org.uk/~.

Keary, Alphonsus, and Walsh (2014). "How affective computing could complement and advance the quantified self." Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on. IEEE

Merriam-Webster.com (2011). "Agitation." http://www.merriam-webster.com (8 May 2011).

Picard (1997). *Affective Computing*. Cambridge: Massachusetts Institute of Technology.

Pierce, Cahill and O'Shea (2014). "Prevalence and Projections of Dementia in Ireland 2011-2046". s.l. : NUIG

WHO (2014). World Health Organization. "Dementia Fact sheet N°362". 28th November 2014.

FindR-TB: A cloud–based tool for antibiotic resistance prediction in *Mycobacterium tuberculosis*

Paul Walsh, Micheál Mac Aogáin, Brendan Lawlor . . . .
Department of Computer Science
CIT – Cork Institute of Technology, Ireland

Department of Clinical Microbiology, School of Medicine,
TCD – Trinity College Dublin.
e-mail: m.macaogain@tcd.ie

Keywords: Cloud computing, Microbiology, Antibiotic resistance, Tuberculosis

The emergence of affordable DNA sequencing technology holds promise for infectious disease diagnostics. Access to whole-genome sequence (WGS) data of pathogenic micro-organisms is now within reaching distance of hospital diagnostic laboratories and promises to bring about positive disruptive changes in diagnostic microbiology. However, without 'point and click' applications addressing computational bottlenecks in the analysis of complex bacterial DNA sequences, clinical adoption of genome-based molecular diagnostics will not be realised.

Mycobacterium tuberculosis (TB) is a global infectious disease that affects approximately 9 million people every year leading to over a 1 million deaths. An estimated 450,000 TB infections exhibit drug resistance. The basis of drug resistance is underpinned by genetic mutations in the TB genome and is therefore predictable given access genomic sequence data. The application of WGS analysis to TB has gained traction as an aid in the diagnosis of TB promising gains in antibiotic resistance detection times and more accurate disease transmission mapping of infected individuals.

Here we report a new development in our cloud-based bacterial DNA management software; the FindR-TB resistance predictor. This tool is integrated into our cloud-based microbial DNA analysis platform (SimplicityTM) and functions in the prediction of antibiotic resistance in *M. tuberculosis* from in-putted user sequence data. In an initial characterisation of our tool we have tested it on genomic data from 100 *M. tuberculosis* isolate genomes allowing us to document the sensitivity and specificity of the tool by using WGS data from phenotyped strains with characterised antibiotic sensitivities.

Diagnostic test statistics were generated for 100 strains allowing us to gauge the potential of this tool as a diagnostic aid for clinical application; i.e. prediction of antibiotic resistance profiles of TB strains and subsequent guidance of therapeutic choices. Resistance to first line antibiotic agents for the treatment of TB including isoniazid, rifampicin, ethambutol and streptomycin was detected with a sensitivity of 91%, 98%, 96% and 100% respectively, and corresponding specificity of 97%, 90%, 73% and 100%. The first line agent pyrazinamide had a significantly lower sensitivity of 34% with a specificity of 100%. These results highlight the ability of our *in silico* predictions to be comparible with currently available diagnostic tests for antibiotic resistant TB while highlighting areas of discord where further characterisation of resistance mechanisms is required.

We plan to test our FindR-TB on a broader collation of TB isolates in order to hone the diagnostic test statistics and improve results. Future tools are currently in development for other pathogenic bacteria including *Clostridium difficile* and *Staphylococcus aureus.* Ultimately these tools will become cornerstones of the clinical microbiology lab as WGS technologies become increasingly integrated into the laboratory work-flows of modern medical diagnostics.

Towards lifecycle management of clinical records in health care environments

Ulrich Kowohl [1], Felix Engel [2], Paul Walsh [3], Alfie Keary [3], Michael Fuchs [4] and Matthias Hemmje [2]

[1] FernUniversität in Hagen, Hagen, Germany

[2] Research Institute for Telecommunication and Cooperation, Dortmund, Germany

[3] Cork Institute of Technology, Cork, Ireland

[4] Wilhelm Büchner University of Applied Sciences, Pfungstadt, Germany

Ulrich.Kowohl@Fernuni-Hagen.de, {Fengel,Mhemmje}@FTK.de, Paul.Walsh@cit.ie, mfuchs@ftk.de

Keywords

SenseCare, Patient Information Lifecycle Management, Patient Data Management, Assisted Living Sensory Data Fusion, Patient Data Management System, Open Archive Standard, Archive Centric Information Lifecycle

Abstract

Sensor Enabled Affective Computing for Enhancing Medical Care (SenseCare) is a project to reduce clinical treatment costs and to enhance life quality of patients with Dementia related disease like Alzheimer. This paper defines usage scenarios of SenseCare related to clinical record data management, outlines current challenges and goals in this field. We will introduce our plan related to this area and present technologies from the Product Data and Product Lifecycle Management field we intend to port to the SenseCare project as well as defining the overall plan to reach our defined goals within the SenseCare project.

1 Introduction

"Affective computing" (Picard et al. (1997)) is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects. AC systems should interpret the emotional state of humans and adapt its behaviour to them, giving an appropriate response for those emotions, or for example alerting carers to the needs of patients who are unable to communicate their needs. In medicine, emotion (human affects) is central to the work carried out by healthcare professionals on a daily basis. It is proposed that "understanding, reflecting on, engaging with and expressing emotion(s) are important aspects of medical practice" and patient care services today (and into the future).

Sensor Enabled Affective Computing for Enhancing Medical Care (SenseCare) is an EC co funded project, which aims to provide a new AC platform based on an Information and Knowledge Ecosystem (RISE, 2015) providing software services applied to the dementia care and connected health domain where enormous potentials and opportunities exist in relation to providing intelligence and assistance to medical professionals, care givers and patients on cognitive states and overall holistic well-being. The SenseCare Ecosystem solution will integrate data streams from multiple sensors (for example: video frames for facial recognition of emotional states, sensory wearables for physiological emotion analytics, and other) and will fuse these streams to provide a global assessment that includes objective levels of emotional insight, well-being and cognitive state. The potential exists to integrate this holistic assessment data into multiple innovative applications

across connected healthcare and various other inter-related and independent domains. In this paper, we now want to address mainly two challenges within the scope of the SenseCare project:

- How can sensory and fused sensory data that come into existence along a patient treatment be systematically managed, processed, analysed, annotated, packaged, and archived to enable its comprehensive long-term preservation, access, and re-use?
- How to represent and assemble provenance information in the sense of temporal bounded treatment data to enable traceability of treatment progress or traceability of decisions made?

Hence, we will present a novel conceptual design for a Person Data Management System (PDMS) which manages and secure patient data across an archive centric Information Lifecycle and at the same time aims at supporting access and reuse of this data. Next, we will discuss several usage scenarios we consider in the SenseCare project. On this basis in section three we will present the current state of the art within the relevant scientific fields. In section four we than sketch a model of a novel PDMS, addressing the above challenges and in section five we define the next to-do's and steps to achieve our goals in section five.

2 Application Scenarios

Dementia is a long-term cognitive and psychologically debilitating disease. It leads to decreasing mental, emotional and social abilities. It reduces cognitive, brain and communication abilities. It is a syndrome which results from several conditions like Alzheimer, Depression, Lewy Dementia or vascular Dementia. Treatment of Dementia is being done with several therapy methods: drugs, brain training, psychological therapies (Cayton et al. 20013). This takes place in care giving environments and last for considerable time. During this, data will be generated: sensory data, clinical data, care plan/execution data and communications. Data underlies a lifecycle that needs to be managed, also in compliance with local regulatory and laws like privacy, minimum and maximal storage durations. In SenseCare we also consider that sensory data will be fused with other multi-sensory data and then analysed, to improve life and care quality for patients living at home with care assistance. Within this overall scope, we mainly address three use cases: ***The Assisted Living Scenario*** deals with home based patients and their daily condition. With monitoring of health conditions, care giving practice can be fine-tuned. We try to support them with an indicator about the patient's condition so that a caregiver can prepare and plan. If caregivers have the possibility to know about a patient's condition upfront, they can already prepare themselves to give more intense or – based on the indication – less intense treatment or alert medical support if there is an indication of a serious health incident. Collecting indicators also enables trend analysis of the mental state of the patient. Also permanent monitoring is part of this use case. Permanent monitoring enables the detection of health incidents so that pre-alerts of medical units can reduce the complexity of possible incidents. Assisted living defines a monitoring environment at the patient's home. The ***Emotional monitoring*** during medical treatment enables clinic personnel to lessen its false/ positive rate during treatment sessions and is described as scenario 2. By having emotional monitoring in place during treatment sessions the therapist is able to judge a patient's behaviour in an improved way and conduct better and safer diagnosis. In addition to successful tracking of dedicated therapy sessions the reaction of the patient related to therapy personal can also be tracked. Since therapy success is also a matter of the interrelations between patient and doctor, we might identify bad relationships and avoid minor to major issues with exchanging therapist/therapy as applicable. ***Shared care giving Scenario*** – e.g. by family members. During treatment processes several groups of persons and actions are generating data. To enable caregivers to have a clear picture of the

patient's condition and its history, data needs to be archived in searchable accessible and interpretable way. The goal is that the collection of information enables stakeholders to provide better medical treatment and diagnostic results. Possible incidents should be identified upfront which makes future diagnosis easier for the doctor since he has some data to work with.

3 State oft he Art

Regnier et al. defines *Assisted Living* as a special combination of housing, supportive services, personalized assistance and healthcare designed to respond to the individual needs of those who require help with activities of daily living and instrumental activities of daily living (Regnier et. al. 2002). In our SenseCare context Assisted Living, Emotional Monitoring and Shared care giving collaboration can be seen as data driven usage scenarios: Each scenario produces data and information and introduces the need to manage and archive them for later reuse. Data management approaches have been already addressed in industrial areas, one of them is Product Data Management (PDM).

PDM, a software oriented solution to manage all relevant information belonging to products generated during the lifetime of the product, aims for re-use of data and for lowering the search time for data during (product) development, production and quality phase. PDM like Parametric Technology's Windchill are currently used by companies in the manufacturing industry like HARMAN Kardon (Ambaruch et al. 2014). PDM systems lower the communication overhead between departments and integrate data streams from various sources into a common data structure with automation. Since PDM deals with products, product data is stored into a data container which encapsulates the namespace of the product. Since product data evolves during time, PDM introduces the possibility to save the current state of the product data container so that it is possible to revisit the status also at a later stage when data has evolved again. This is called baselining. Since enhanced searchability of data is a key aspect of PDM, better searchability of information – as one key aspect of the usage scenarios - can be reached by common generic structures of products having the same type which is one PDM approach. Instances of generic structures are called Bill of Information (BOI, Tekale et al. 2012). The Bill of information is one concept addressed in the PDM area. The BOI has two parts: Generic structure – Upper BOI – and product data: Lower BOI. While upper BOI defines the generic common data structure, the lower structure is defined by actual product data. Lower BOI data is loosely linked into upper BOI so that data from the lower BOI is grouped by construction parts and sliced due to its aspects like mechanical CAD, electronically CAD, software, quality assurance, communications and regulatory.

Product Engineering takes place with several departments and stakeholders. Those Stakeholders have to collaborate together. Workflows consisting of work packages can reduce communication overhead. Work packages belong to tasks which have to be executed on product data structures within the lower BOI. Workflows automate communication, e.g. between different departments. For example: when manufacturing some electronic control circuit, a workflow is established to check, if the electronic circuit – designed by the electronics department – actually fits into its designated case – which was designed by the mechanical CAD department. Tasks on data can be routed through several departments and can end up with a managerial approval which changes states of related design documents to be ready for the next development step or production ready.

As shown, PDM can introduce approaches to enhance searchability of data and to collaborate on data through a lifecycle based workflow. SenseCare has the need to establish such approaches in a Patient Data Management System. **Patient data management systems** (PDMS) are software

systems that integrate administrative functions and support clinical decision-making processes. They are responsible for managing patient data within hospitals and support hospital processes. PDMS also support hospitals to meet their requirements of the contracting environment e.g. regarding financial aspects. They further support the handling of electronic patient records (EPR). While one EPR describes the record of a singular time boxed care process executed by a single institution (Fretscher et al. 2001, Mitey et al. 2001). **Archival of research data** is the key mechanism to successfully support later access to research data. The Reference Model for an *Open Archival Information Systems* (OAIS, 2009) is a frequently applied framework of common terms and concepts for an archival information system that suggests the usage of so called Information Packages for describing an archived object.

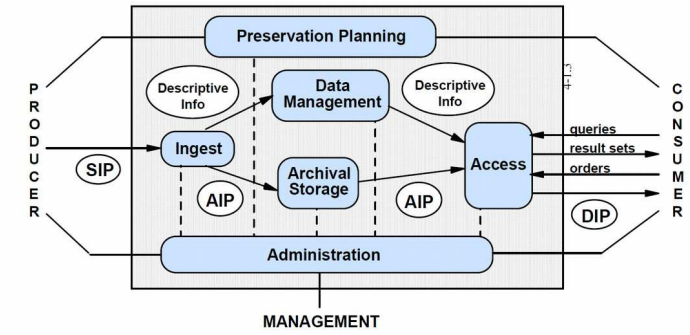


Figure 1. OAIS functional entities (cf. CCSDS, 2009)

The Figure 1 provides an overview of all involved actors, components and processes, that are part of an Information Package archival. Our work in the SHAMAN research project (Engel et al. 2009) outlined, that from the perspective of long-term preservation, digital encoded objects undergo an archive-centric information life cycle with the phases Creation, Assembling, Archiving, Adoption, and Reuse (Brooks et al. 2010):

- **Creation**: new information comes into existence.
- **Assembly**: denotes the appraisal of objects relevant for archival and all processing and enrichment for compiling the complete information set to be sent into the future, meeting the presumed needs of the Designated Community. Assembly requires in-depth knowledge about the Designated Community in order to determine objects relevant for long-term preservation together with information about the object required for identification and reuse some time later in the future.
- **Archival**: addresses the life-time of the object inside the archive.
- **Adoption**: encompasses all processes by which accessed archival packages are unpacked, examined, adapted, transformed, integrated and displayed to be usable and understandable for the consumer. This includes also emulation activities if needed. The adoption phase might be regarded as a mediation phase, comprising transformations, aggregations, contextualisation, and other processes required for re-purposing data.
- **Reuse**: means the exploitation of information by the consumer. In particular, reuse may be for purposes other than those for which the Digital Object was originally created. Reuse of Digital Objects can lead to the Creation of other, novel Digital Objects. Reuse also may instigate the addition or updating of metadata about the Digital Object held in the archive. For example, annotation changes informational content and affects the relationships existing between the Object and other Digital Objects.

Within SHAMAN and the SCIDIP-ES project (Crompton et al. 2014) we envisioned and partly implemented a Packaging Service for packaging digital encoded objects for long term archival under consideration of information lifecycle that come into existence during the single phases of the object lifecycle, in conformance to the OAIS Information Model specification. The Packaging Service is software that specifically supports the creation and management of Information Packages. It has been developed and enhanced during several EC co funded research projects. Hence, the *Packaging Service* has been already applied to different user communities as in the context of memory institutions within the *SHAMAN* project, besides the earth science community (SCIDIP-ES). Actually, the Service fulfils functions for handling:

- **Single file AIP**: the aim is to create a single file which can be sent to a user who cannot access the information in a repository remotely.
- **Distributed packaging – simple Archival Information Unit (AIU)**: A user wishes to collect together all the information needed for the AIP of a single dataset (i.e. an AIU).
- **Distributed packaging – Archival Information Collection (AIC)**: Repository manager needs to create AIPs for a large collection of data of say 100000 individual data files.

At its heart the management of AIPs, AIUs and AICs are implemented through the ability to create/parse a so called *Information Package manifest file* that is constructed and populated with concepts and instances in conformance to the OAIS Information Model, providing the information and/or libraries necessary for de-serializing the archived digital encoded object. Since an archived digital encoded object could be complex and interlinked with (external) resources, archiving them requires a structured and machine processable serialization mechanism. Such mechanisms should be able to capture and preserve a digital object composition and all relevant relationships. Hence, the *Packaging Service manifest file* is serialized through the application of the *OAI-ORE* specification (Nelson et al. 2008) as the object model to represent all constituents and their interrelation. The *OAI-ORE* builds on the principles of the RDF specification as defined by the W3C. It defines standards for the description of aggregations of web resources and the digital objects of which they are composed.

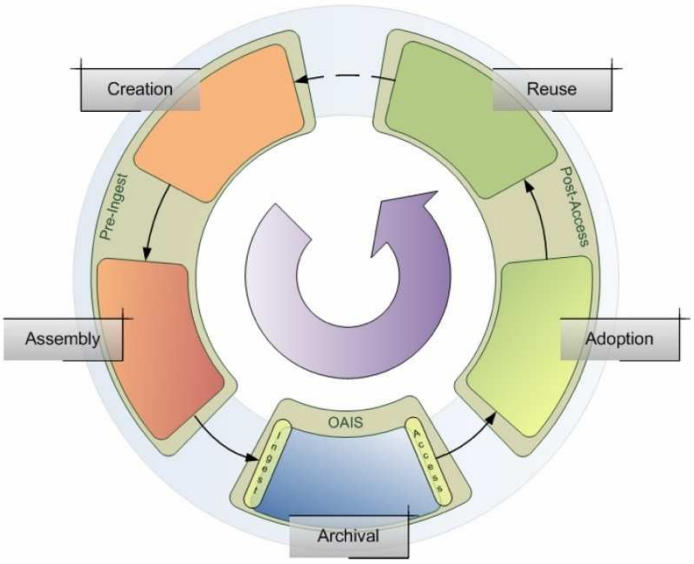


Figure 2. Information lifecycle phases (Engel et al. 2009)

4 Solution Draft

The outlined use cases showcase that from the perspective of the archive centric life cycle various information comes into existence during a patient treatment that are of potential use for later exploitation, to increase a patient's wellbeing. We therefore propose the introduction of a *Patient Lifecycle Management system* (PLMS) that support lifecycles of information. PLMS will use approaches from PDM/PLM and PDMS systems. While existing PDMS are specialized EPR record archives, our PLMS system extends the functionality of existing PDMS in a way that it supports the archive centric information lifecycles as well as it integrates process based collaboration between different stakeholders which can belong to different institutions over collaboration lifecycles. Our PLMS consists of several information containers whith each container holding all relevant data belonging to a patient. Each container will have a lower and an upper BOI, while the upper BOI can be enhanced with different additional aspect oriented generic structures. This upper BOI approach is based on the assumption, that it is possible to create a generic structure describing patient history, therapy approaches etc. related to dementia. Figure 3 displays one example instance of our PLMS with one patient.

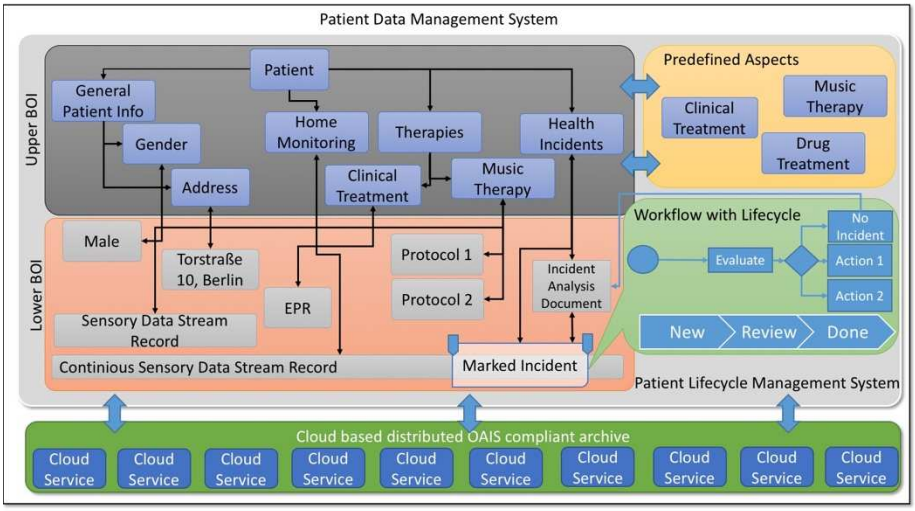


Figure 3. Novell Approach with one Patient Data Container

As displayed, PLMS and the OAIS archive are decoupled in two layers. Following the archive centric lifecycle of information, data will be created during the creation phase and starts with the video recordings done by data producers. In the *Assembly phase* the video needs to be enhanced to represent an OAIS Submission Information Package or SIP by contextualization with emotion trajectories – output of an emotion detection algorithm mapped to a generic emotion taxonomy – and additional meta data e.g. the applied analysis algorithms, location, dates or else. Emotion trajectory (Figure 4) is a time ordered set in n-dimensional space where each dimension represents an emotion from a generic emotion-taxonomy which enables us later on to exchange emotion detection algorithms.

$$\begin{pmatrix} e_1 \\ \dots \\ e_n \\ t_0 \end{pmatrix}, \dots, \begin{pmatrix} e_1 \\ \dots \\ e_n \\ t_m \end{pmatrix}$$

Figure 4. Emotion Trajectory

This assembly task and the representation of the interrelation of the various involved resources could hence be implemented by means of the Packaging Toolkit. In essence the Toolkit produces an information container that holds all relevant information for archival and later access. After ingestion, in the *Archival phase* all information belonging to a patient is than archived as an Archival Information Package (AIP). In parallel to archiving the AIP, the AIP needs to be linked to the dedicated substructures of the upper BOI it belongs to. Searchability of information is then guaranteed by the upper BOI. Interpreting such data generates the need to collapse video annotations to a simple, understandable information about the health status of the patient based on data consumers and will consists of views as part of the Adoption phase. Collaboration means reusing data to a specific need and is part of the reuse phase. Collaboration between stakeholders will be ensured by workflows. Workflows can be executed on a selected subset of BOI which then will be handled as a temporary information container. Outcome of tasks in workflows can be additional data, where the Information Lifecycle starts for the information again. The example in Figure 3 shows exactly this case. A video stream has been archived through the creation phase. Parts of this archived information package have been marked as interesting for further investigation (Marked Incident). The specific data has been extracted as an additional information object with review workflow task (IORW) with a PLMS collaboration lifecycle: New, Review, Done. Once the information has been extracted, the IORW is in status New. A doctor or another information consumer needs to review it. After reviewing the workflow produces a diagnosis document as an outcome which needs to be archived through the archival phases as well.

5 Methodology

Figure 3 shows the overall approach based on an example novel with one patient information container which is represented as grey structure consisting of Upper BOI and Lower BOI. Additionally, to current PDMS functionality, the collaboration workflow as well as lifecycle model is visible by having an incident detected which needs to be investigated by a doctor. An incident has 3 lifecycle states (New, Review, Done) and results in an incident analysis document being added to patient information. For the create phase we will establish an upload mechanism for videos in the cloud based education portal. For supporting assembly phase there is the need to project output of emotion annotation algorithms to a n+1 dimensional trajectory while n is the size of SenseCare's own defined emotion taxonomy, the additional n+1 dimension represents time. A mapping from algorithm taxonomies to generic ones needs to be developed. This step will be embedded in the use of the Packaging Toolkit. In section 2 we introduced scenarios that are quite generic. In a first step we will concretize those and conduct user stories. Based on defined user stories we will identify stakeholders, resources and processes. With detailed definition of user stories, stakeholders, involved resources and processes we will assemble test data and technical specifications of involved medical analysis tools from our SenseCare partners. This will help us to get a deeper understanding of the medical domain and necessary solutions. Having gained this deep knowledge about user stories, test data, processes and stakeholders we will create a set of use cases for each user story including functional and technical requirements. In parallel we will create test cases out of defined user stories and collected functional and technical requirements. This means that test cases will completely cover usage scenarios and can be used for validating the developed approach during solution development. Design scenarios have to be evaluated during the archival phase to have a clean and well-structured AIP. Those considerations will be done in the archival strategy plan which has to be developed. Data management planning activities such as stakeholder definition, definition of output scenarios and access rights, data views and sharing policies have to be identified for the adoption phase. Establishing the Patient Information Lifecycle Management

system will be handled in planning the reuse phase. Here scenarios, data workflows as well as additional stakeholders have to be identified to support usage scenario three and four.

Additionally, there needs to be the definition of a set of high level upper - BOI and related workflows. This BOI will lead to already fixed information structure to enable reuse scenarios like usage scenarios three and four. Having upper – BOI defined there will be the need to cut this resulting graph into pieces – so called views – according to their dedicated aspects. Those views need to be defined according to additional requirements of sub use cases and need to be reflected in the Dissemination package.

6 Summary and Outlook

This paper gave a brief overview about the SenseCare project and described SenseCare usage scenarios from a data management perspective. Next to standard data management planning activities we described the further need of data management approaches for:

- Relating EPR, collaboration, sensory data, fused sensory data as well as other clinical data
- Enhancing the ability to find already archived data and to speedup searching processes
- Introducing further approaches from PDM like workflows and workflow based lifecycles to optimize collaboration between stakeholders

For achieving this, we outlined that we will port approaches based in the Product Data Management and Product Lifecycle Management area to the domain of clinical treatment and management of patient data in clinical treatment usage scenarios. The purpose and the status of the project related to data management has been shown and the next steps like definition of detailed user stories has been presented.

7 Acknowledgment


This publication has been produced in the context of the SenseCare project. This project has received funding from the European Union's H2020 Programme under grant agreement No 690862. 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.

8 References

Ambaruch, B., Shandwick, W., 2016, PTC's PLM Momentum Accelerates as Automotive Industry Evolves, PTC, BusinessWire. Available at: http://files.shareholder.com/downloads/PMTC/0x0x524572/4a74a130-3109-42b8-9446-a0abb97b5e58/PMTC_News_2011_2_14_General.pdf, Accessed 2016-03-01

Brocks, H., Kranstedt, A., Jäschke , G., & Hemmje, M., 2010, Modelling context for digital preservation. *Stud. Comput. Intell.*, 197–226

CCSDS, 2009, Reference model for an open archival in-formation system (OAIS), ISO 17421

Cayton, H., Nori G., and Warner J., 2002 *Dementia: Alzheimer's and Other Dementias: The'at Your Fingertips' Guide: The Fully Updated and Comprehensive Reference Book for Alzheimer's and Other Forms of Dementia*. Class Publishing Ltd

Cromton, S., Giaretta, D., Matthews, B., Engel, F., Brocks, H., Shaon, A., & Marelli, F, 2014, A Sustainable Data Preservation Infrastructure to Support OAIS Conformant Archives. *APA/C-DAC International Conference on Digital Preservation and Development of Trusted Digital Repositories*. New Dehli, Inda

Engel, F., Klas, C., Brocks, H., Kranstedt, A., Jäschke, G., & Hemmje, M., 2009, Towards supporting context-oriented information retrieval in a scientific archive based information lifecycle. Proceedings of Cultural Heritage online. Empowering users: an active role for user communities, (pp. 135-140). Florence, Italy.

Fretschner, R., Bleicher, W., Heininger, A. and Unertl, K., 2001 Patient data management systems in critical care. *Journal of the American Society of Nephrology*, 12(suppl 1), S83-S86.

Lagoze, C., Van de Sompel, H., Nelson, M.L., Warner, S., Sanderson, R. and Johnston, P., 2008. Object re-use & exchange: A resource-centric approach. *arXiv preprint arXiv*:0804.2273

Mitev, N. and Kerkham, S., 2001. Organizational and implementation issues of patient data management systems in an intensive care unit. *Journal of Organizational and End User Computing*, 13(3), p.20

Nawroth, C., Schmedding, M., Brocks, H., Kaufmann, M., Fuchs, M., & Hemmje, M. (2015). Towards Cloud-Based Knowledge Capturing Based on Natural Language Processing. *Procedia Computer Science*, volume 68,pp 206-216

Pichard, R., Affective Computing (1997), MIT Press Cambridge, MA, USA

Reasearch and Innovation Staff Exchange (RISE), Sensor Enabled Affective Computing for Enhancing Medical Care, Call H2020-MSCA-RISE-2015, 2015

Regnier, V., 2003 *Design for assisted living: Guidelines for housing the physically and mentally frail*. John Wiley & Sons.

Tao, J. and Tan, T (2005), Affective Computing: A Review, Lecture Notes in Computer Science, volume. 3784, pp 981-995

Tekale S. K., Nandedkar V. M., 2012, PLM Implementation with Creation of Multiple Item using SOA Technology in Team center. *IOSR Journal of Mechanical and Civil Engineering* (IOSR-JMCE), ISSN(e) : 2278-1684, ISSN(p) : 2320–334X, PP : 62-66, 2012

Eye-Tracking in Computer-Based Simulation in Healthcare Training

Jonathan Currie, Raymond R. Bond, Paul McCullagh, Pauline Black, Dewar Finlay
School of Computing and Mathematics, Jordanstown Campus
Computer Science Research Institute
Ulster University, United Kingdom
e-mail: currie-j@email.ulster.ac.uk

Keywords: Eye-Tracking, Visual Attention, Simulation-Based Training, Task Performance

Introduction

Patient safety is a critical area of concern within healthcare and medical errors are a well-known problem that can have fatal ramifications (Kohn et al. 2000). Lack of knowledge and skill with clinical tasks and procedures, as well as decision-making can be significant factors with many of the errors that are reported in healthcare (Zhang et al. 2002). Many healthcare tasks can be simulated using computer and web technology for training purposes and provide trainees (students and practicing) with a way to improve or maintain their knowledge and skills (Persson et al. 2014; Cant & Cooper 2014). The concept of visual attention during a task has been tested in medical and healthcare task studies (O'Meara et al. 2015; Zheng et al. 2011; Breen et al. 2014) with an aim of finding discriminative differences between competency levels. The study of this previous work led us to hypothesise that eye tracking metrics exclusively have a relationship with specific task performance and can discriminate between performance level. We found the duty of patient monitoring with interpreting vital signs monitors in nursing earmarked in the literature for improvement in available simulation-based training. We sought to use eye-tracking with the task of nurses interpreting simulated patient vital signs from a monitor. The objective was to determine if eye-tracking technology can be used to develop biometrics for automatically classifying the performance of nurses whilst they interact with computer-based simulations.

Methods

For the study a total of 47 nurses were recruited, with 36 nursing students (TG - Training Group) and 11 coronary care nurses (QG - Qualified Group). Each nurse interpreted five simulated vital signs scenarios whilst 'thinking-aloud'. We recorded the participant's visual attention (eye tracking metrics [ETMs]), verbalisation, confidence level (1-10, 10=most confident). Interpretations were given a score out of ten to measure performance. Data analysis was used to find patterns between the ETMs and performance. Multiple linear regression was used to predict performance (score) using ETMs.

The five scenarios were designed by an expert nurse with validation was provided by three colleagues of similar expertise. The scenarios were designed to be within the expected ability of the average undergraduate student but also with assessment criteria that would likely provide fully trained and qualified nurses a higher score.

The criteria assessed participants' verbal responses at three levels and scores were allocated according to:

Basic performance: identification of abnormalities in the presented vital signs.

Mid-range performance: identification of why the abnormalities occurred based on their knowledge and understanding of the presenting condition outlined in the case scenarios.

High-range performance: identification of and decision-making with regard to the immediate interventions required to stabilise the patient.

Performance scores were then put into classes according to expert advice:

- 0-5 = low
- 6-7 = medium
- 8-10 = high

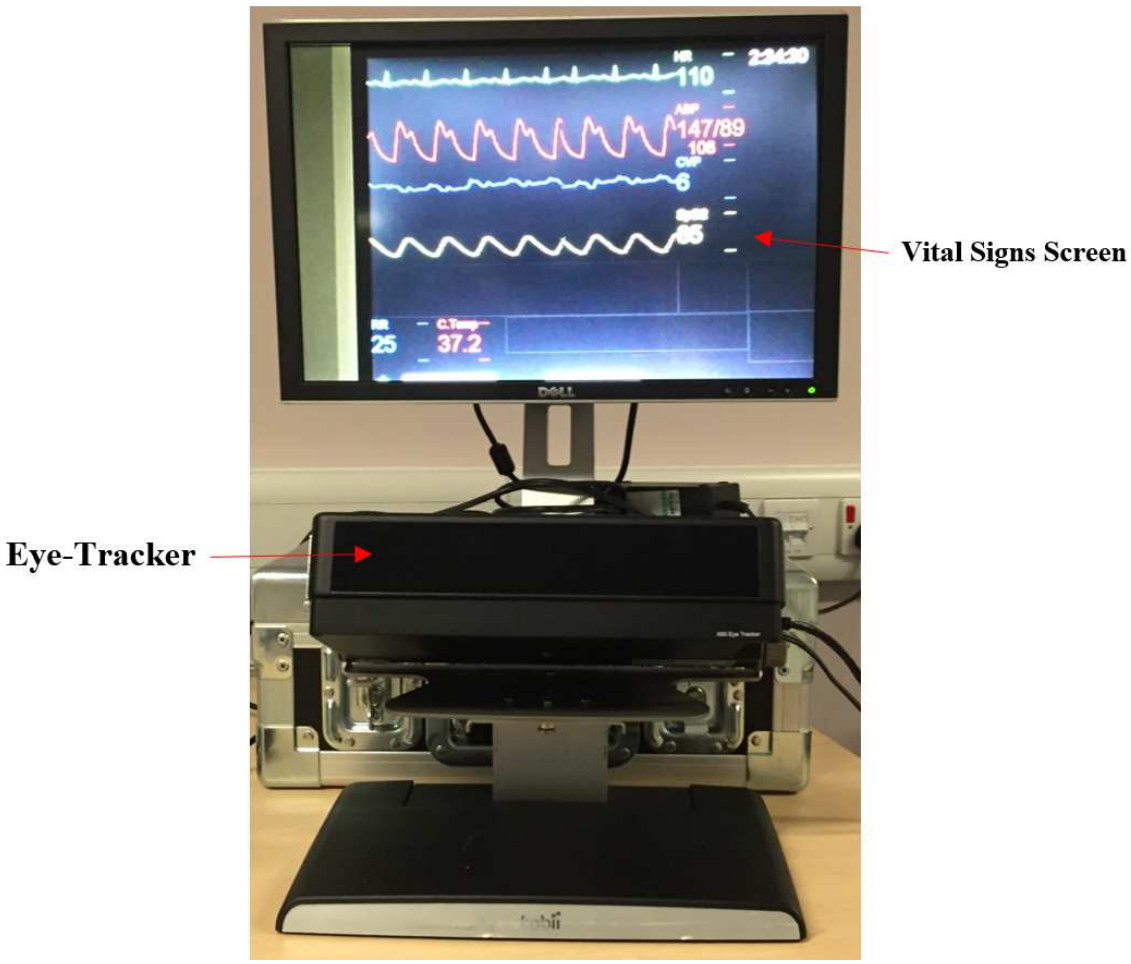


Figure 1. Participant Point of View During Interpretations

Figure 2. Simulated Vital Signs Screen with Eye Tracking Areas of Interest with Variable Names Highlighted

Results

The QG scored higher than the TG (6.85±1.5 vs. 4.59±1.61, p=<0.0001) and reported greater confidence (7.51±1.2 vs. 5.79±1.39, p=<0.0001).

Class	TG	QG	All
Low (0-5)	129 (72%)	10 (18%)	139 (59%)
Medium (6-7)	45 (25%)	25 (45%)	70 (30%)
High (8-10)	6 (3%)	20 (36%)	26 (11%)

Table 1. Interpretation Performance Results in Classes (Count and %)

20 ETMs discriminated between classes of performance (low, medium, high). Regression using ETMs (62 independent variables) was shown to adequately predict score (adjusted R2=0.80, p=<0.0001).

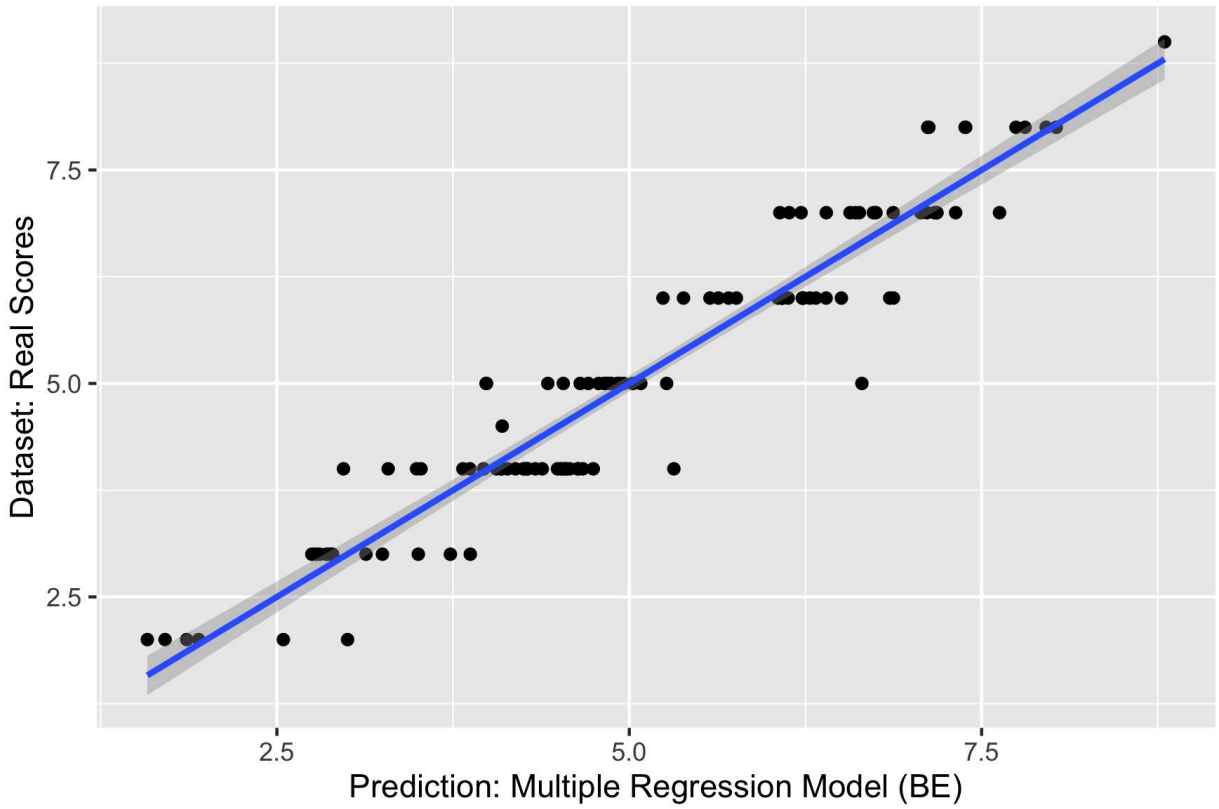


Figure 3. Participant Scores Vs Predicted Scores Using Subset of Eye Tracking Metrics As Predictors

Conclusions and Future Work

Thedata collected, specifically the data analysis of ETMs,has shown that visual attention and the level ofperformance for this specific task (measured byperformance score) are not independent of each other. Put differently, ETMs exclusively could be used topredict a person's performance when reading vital signs. Explicitly, we can see that specific ETMs are statisticallysignificant in discriminating between classes ofperformances for interpreting vital signs. With furtherrecruitment of participants (especially high classperformances), we could potentially reveal more ETMsthat are discriminatory between low, medium and highclass performance. Further research, including statisticaltechniques like PCA are required to refine the regressionmodels and the optimal level of accuracy for predictingperformance using the ETMs. At present, we canconclude that there is a relationship between ETMs and performance that can be seen but it is unclear to what extent ETMs can predict performances.

Future newly designed studies will include:

1. Using eye-tracking during a much more complex simulated procedure, such as cardio angiography. We want to test the assumption (Zheng et al. 2011) that trainees and experts have different patterns of visual attention during surgical tasks that classify them. During this, we will also look at the link between visual attention and attentional capacity (Weaver 1949) of the task performer by provided them with added stimulus/distraction.

2. Other future work will progress the study described above, with eye-tracking to discriminate performance whilst an experimental group is presented a changed layout of on-screen vital signs. This work will look at the influence of information hierarchy on interpretation performance as the experimental group is presented the same scenarios but with the most concerning vitals emphasized on screen (positioning). We wish to test if this provides any advantage to the experimental group in their performance. If so, it could make an argument for automatic prioritizing of vital signs in accordance to the specific patient (through machine learning algorithms) in future equipment to reduce poor interpretations and patient monitoring.

References

Breen, C.J., Bond, R. & Finlay, D., 2014. An evaluation of eye tracking technology in the assessment of 12 lead electrocardiography interpretation. Journal of electrocardiology, 47(6), pp.922–9. Available at: http://www.sciencedirect.com/science/article/pii/S0022073614003148.

Cant, R.P. & Cooper, S.J., 2014. Simulation in the Internet age: the place of web-based simulation in nursing education. An integrative review. Nurse education today, 34(12), pp.1435–42. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84906073568&partnerID=tZOtx3y1 [Accessed July 19, 2015].

Kohn, L.T., Corrigan, J.M. & Donaldson, M.S., 2000. To Err is Human: Building a Safer Health System, National Academies Press (US). Available at: http://www.ncbi.nlm.nih.gov/pubmed/25077248 [Accessed July 31, 2016].

O'Meara, P. et al., 2015. Developing situation awareness amongst nursing and paramedicine students utilizing eye tracking technology and video debriefing techniques: A proof of concept paper. International emergency nursing, 23(2), pp.94–9. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84926483577&partnerID=tZOtx3y1.

Persson, J. et al., 2014. Evaluating interactive computer-based scenarios designed for learning medical technology. Nurse Education in Practice, 14(6), pp.579–585. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84918813292&partnerID=40&md5=ab8235ae36ae3a782c634b9afbf51abb.

Weaver, W., 1949. THE MATHEMATICS is closely linked with the concept of information. , 181(1), pp.11–15.

Zhang, J., Patel, V.L. & Johnson, T.R., 2002. Medical error: is the solution medical or cognitive? Journal of the American Medical Informatics Association?: JAMIA, 9(6 Suppl), pp.S75-7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=419424&tool=pmcentrez&rendertype=abstract.

Zheng, B. et al., 2011. Surgeon's vigilance in the operating room. The American Journal of Surgery, 201(5), pp.673–677. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0002961011001140.

Chapter 4

Computing

Harassment detection: a benchmark on the #HackHarassment dataset

Alexei Bastidas, Edward Dixon, Chris Loo, John Ryan
Intel
e-mail: edward.dixon@intel.com

Keywords: Machine Learning, Natural Language Processing, Cyberbullying

Introduction

Online harassment has been a problem to a greater or lesser extent since the early days of the internet. Previous work has applied anti-spam techniques like machine-learning based text classification (Reynolds, 2011) to detecting harassing messages. However, existing public datasets are limited in size, with labels of varying quality. The #HackHarassment [1] initiative (an alliance of tech companies and NGOs devoted to fighting bullying on the internet) has begun to address this issue by creating a new dataset superior to its predecssors in terms of both size and quality. As we (#HackHarassment) complete further rounds of labelling, later iterations of this dataset will increase the available samples by at least an order of magnitude, enabling corresponding improvements in the quality of machine learning models for harassment detection. In this paper, we introduce the first models built on the #HackHarassment dataset v1.0 (a new open dataset, which we are delighted to share with any interested researcherss) as a benchmark for future research.

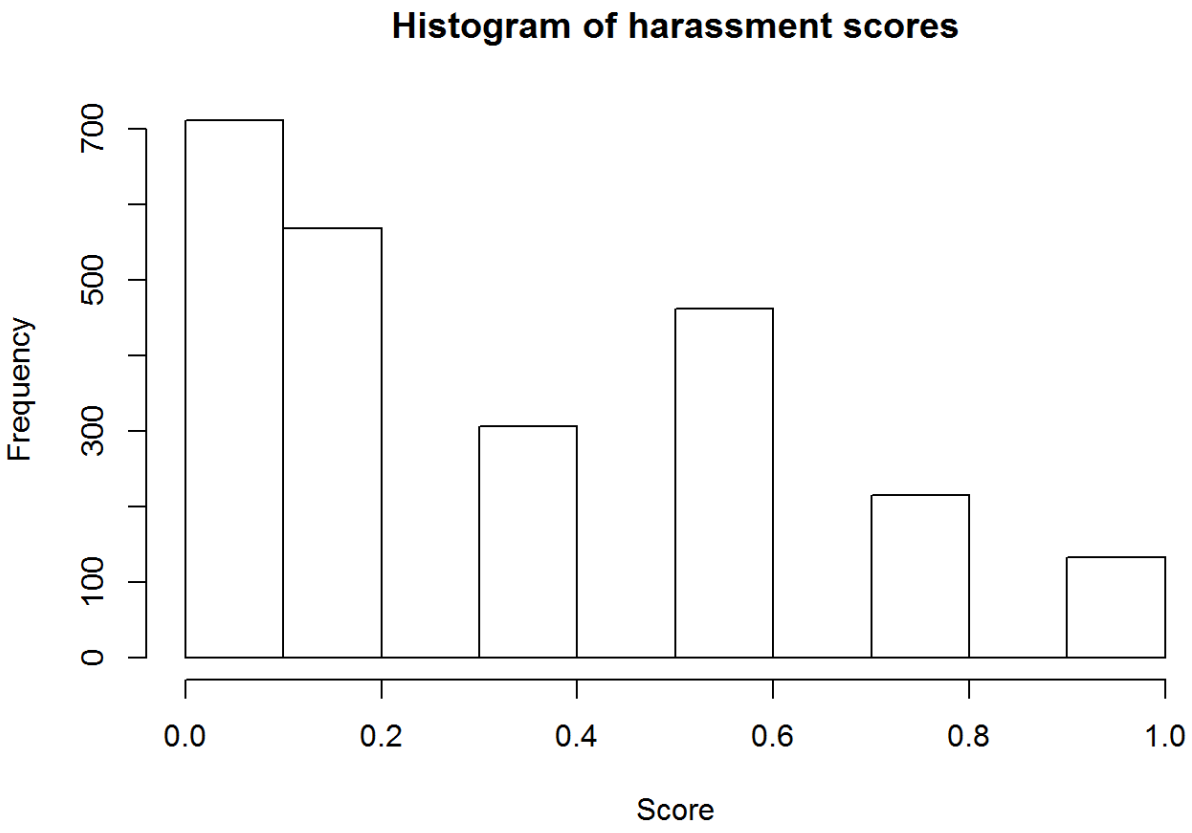
Related Work

Previous work in the area by Bayzik 2011 showed that machine learing and natural language processing could be successfully applied to detect bullying messages on an online forum. However, the same work also made clear that the limiting factor on such models was the availability of a suitable quantity of labeled examples. For example, the Bayzick work relied of a dataset of 2,696 samples, only 196 of which were found to be examples of bullying behaviour. Additionally, this work relied on model types like J48 and JRIP (types of decision tree), and k-nearest neighbours classifiers like IBk, as opposed to popular modern ensemble methods or deep neural-network-based approaches.

Methodology

Our work was carried out using the #HackHarassment Verison 1 dataset, the first iteration of which consists exclusively of Reddit posts. An initially random selection of posts, in which harassing content occured at a rate of between 5% and 7% was culled of benign content using models training on a combination of existing cyberbullying datasets (Reynolds 2001, also “Improved cyberbullying detection through personal profiles). Each post is labelled independently by at least five Intel Security Web Analysts. (a post is considered “bullying” if it labelled as such by 20% or more of the human labelers - as shown in the following histogram, a perfect consensus is relatively rare, and so we rate a post as “harassing” if 20% - 2 of our 5 raters - consider it to be harassing). This is a relatively balanced dataset, with 1,280 non-bullying/harassing posts,, and 1,118 bullying/harassing examples.

1 "Hack Harassment." 2016. 26 Jul. 2016 <http://www.hackharassment.com/>



All pre-processing, training and evaluation was carried out in Python, using the popular SciKit-Learn library [2] (for feature engineering and linear models) in combination with Numpy [3] (for matrix operations), Keras [4] and TensorFlow [5] (for models based on deep neural networks - DNNs).

For the linear models, features were generated by tokenizing the text (breaking it aparting into words), hashing the resulting unigrams, bigrams and trigrams (collectiojns of one, two, or three adjacent words) and computing at TF/IDF for each hashed value. The resulting feature vectors were used to train and test Logistic Regressioin, Support Vector Machine and Gradient Boosted Tree models, with 80% of data used for training and 20% held out for testing (results given are based on the held-out 20%).

For the DNN-based approach, a similar approach was taken to tokenization, both bigram and trigram hashes were computed; these were one-hot encoded, and dense representations of these features were learned during training, as per Joulin 2016.

2 "scikit-learn: machine learning in Python — scikit-learn 0.17.1 ..." 2011. 29 Jul. 2016 <http://scikit-learn.org/>

3 "NumPy — Numpy." 2002. 29 Jul. 2016 <http://www.numpy.org/>

4 "Keras Documentation." 2015. 29 Jul. 2016 <http://keras.io/>

5 "TensorFlow — an Open Source Software Library for Machine ..." 2015. 29 Jul. 2016 <https://www.tensorflow.org/>

The FastText model used is a python implenmentation of the model described in "Bag of Tricks for Efficient Text Classification." [6] . For the text encoding, bigrams and trigrams are used. 20% of the data was held out for testing.

The Recurrent Character Level Neural Network model consists of 2 GRU layers of width 100 followed by a Dense Layer of size 2 with softmax on the output, Between each of the layers batch normailization is performed. The optimiser used was rmsprop. For data preperation each of characters was onehot encoded and each sample was truncated/padded to 500 charcters in length. 20% of the data was held out for testing.

Results

Model	Precision (Harassing)	Recall (Harassing)
Gradient Boosted Trees (Scikit-Learn)	0.80	0.71
Bernoulli Naive Bayes	0.54	0.30
FastText	0.60	0.78
Recurrent Character Level Neural Network	0.71	0.73

Conclusions

We have presented the first results on a new open cyberbullying/harassment dataset. While our models clearly demonstrate a degree of ability to discriminate between the content classes, the achieved precision in particular falls far short of our ambitions for #HackHarassment.

Over the coming months, we'll massively expand the size of our labelled dataset, and review our labelling methodology, anticipating that a larger dataset will facilitate more accurate models. We look forward both to the availability of a larger dataset, and to seeing the development of classifiers that improvement on our work, and welcome partners able to contribute either in terms of expanding the dataset or improving the modelling.

6 "fastText" 2015. 22 Jul. 2016 <https://github.com/sjhddh/fastText>

References

Reynolds, Kelly, April Kontostathis, and Lynne Edwards. "Using machine learning to detect cyberbullying." Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on 18 Dec. 2011: 241-244.

Bayzick, Jennifer, April Kontostathis, and Lynne Edwards. "Detecting the presence of cyberbullying using computer software." (2011): 1-2.

Joulin, Armand et al. "Bag of Tricks for Efficient Text Classification." arXiv preprint arXiv:1607.01759 (2016).

Improved Cyberbullying Detection Through Personal Profiles http://doc.utwente.nl/80761/

Large-Scale Biomedical Data Integration and Data Mining: a Multiplex Network-based Approach

Haiying Wang, Huiru Zheng
School of Computing and Mathematics, Computer Science Research Institute
University of Ulster, N.Ireland, UK
e-mail: {hy.wang, h.zheng}@ulster.ac.uk

Keywords: Data integration, multiplex networks

Progress in medical sciences and molecular biology has led to the accumulation of tremendous amounts of biomedical data. In addition to traditional clinical data such as physical examinations, electrocardiogram (ECG) recordings and medical images, biomolecular data including DNA, RNA, protein sequences, and their two-dimensional (2D) and three-dimensional (3D) structures are being accumulated at an exponential rate in in both quantity and diversity, providing unique opportunities to study biological systems at different levels (Ritchie et al., 2015). For example, as a large-scale, collaborative effort led by the National Institute of Health, The Cancer Genome Atlas (TCGA) has collected massive, high quality information generated from various molecular levels for over 30 types of human cancer offering a valuable resource to accelerate our understanding of the molecular basis of human cancers (Cancer Genome Atlas Research, 2013).

Analysing such volume of biomedical data characterised by the presence of large amounts of high-dimensional data and a variety of fuzzy and noisy forms represents not only an enormous challenge but significant opportunity. Over the past decades, a wide range of computational approaches have been proposed and developed. Examples include iCluster (Shen et al., 2009), Similarity Network Fusion (SNF) (Wang et al., 2015), and novel statistical approaches (Pineda et al., 2015). In this study, an integrative, multiplex network-based approach was introduced to explore heterogeneous biomedical data (Mucha et al., 2010, Wang et al., 2016). For each given dataset, a similarity network will be constructed, in which each node corresponds to a subject and each edge represents the similarity between a pair of subjects derived from the given dataset. By introducing a coupling between networks to account for potential correlations between different data sources, a new multiplex network-based framework for integrative analysis of heterogeneous datasets, which has potential to extract homogeneous patterns shared by datasets, will be developed and implemented.

The system has been tested on the identification of the subtypes of glioblastoma multiforme (GBM) and breast invasive carcinoma from three omics data, i.e. i.e. mRNA expression, DNA methylationa, and miRNA expression. It has been shown that the subtypes identified are closely correlated with clinical variables as shown in Fig. 1 and a high level of concordance indicated by the value of Normalized Mutual Information (NMI) has been achieved in comparisons to state-of-the-art techniques (NMI > 0.8).

The preliminary results (Wang et al., 2016) have demonstrated that the proposed method represents a useful alternative network-based solution to large scale, integrative analysis of heterogeneous biomedical data and a significant step forward to the methods already in use in which each type of data are treated independently. It has several advantages. For example, the framework has the ability to correlate and integrate multiple data levels in a holistic manner to facilitate our understanding of the pathogenesis of disease. It provides a flexible platform to integrate different types of patient data, potentially from multiple sources, allowing discovering complex disease patterns with multiple facets. Currently we are applying the techniques to extract patterns associated with Alzheimer's disease based on integration of longitudinal clinical data and MRI images.

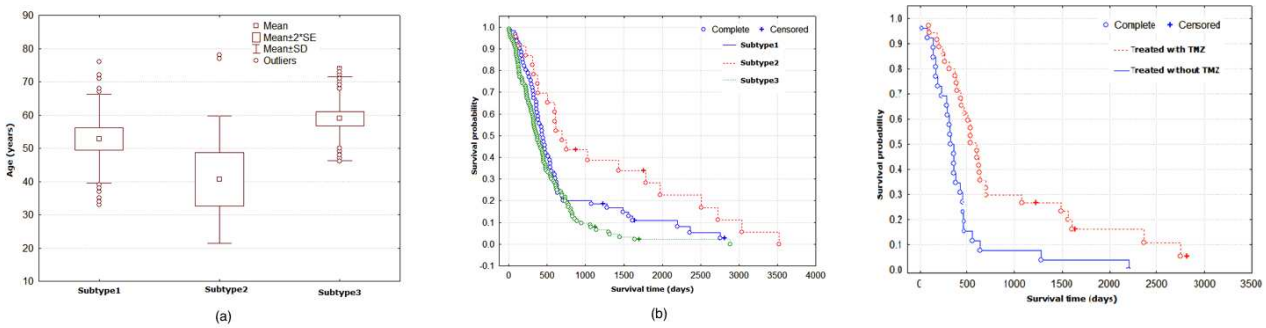


Figure 1. The correlation with clinical variables: (a) A statistically significant difference in terms of the average age was observed across 3 GBM subtypes (ANOVA test, $p < 0.0001$) with Subtype2 is closely associated with younger patients; (b) Survival times are significantly different among three GBM subtypes with patients in Subtype 2 having a more favorable prognosis; and (c) Survival analysis of GBM patients for treatments with temozolomide (TMZ) in Subtype1. A significantly increased survival time was observed (Cox log-rank test, $p < 0.005$).

References

Cancer Genome Atlas Research Network (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, 2013, 497(7447), pp.67-73

Mucha, P., Richardson, T., Macon, K., Porter, M., and Onnela, J. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, vol. 208, pp. 876-878.

Pineda, S., Real, F.X., Kogevinas, M., Carrato, A., Chanock, S.J., Malats, N., et al. (2015). Integration Analysis of Three *Omics* Data Using Penalized Regression Methods: An Application to Bladder Cancer. *PLoS Genet*, 11(12): e1005689.

Ritchie, M. D., Holzinger, E. R., Li, Pendergrass, S. R., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* 16, pp.85–97

Shen, R., Olshen, A., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 2009, 25, pp. 2906–2912.

Wang, B., Mezlini, A.M., Demir, F., Flume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, 11, 333–337.

Wang, H., Zheng, H., Wang, J., Wang, C., and Wu, F. (2016) Integrating omics data with a multiplex network-based approach for the identification of cancer subtypes. *IEEE Transactions on Nanobioscience*, in press.

Towards a Cross Industry Standard Process to support Big Data Applications in Virtual Research Environments

K. Berwind [1], M. Bornschlegl [1], M. Hemmje [1], M. Kaufmann [2]

[1] Faculty of Mathematics and Computer Science, University of Hagen, Germany

[2] Engineering & Architecture, Lucerne University of Applied Sciences and Arts

{kevin.berwind, marco-xaver.bornschlegl, matthias.hemmje} @fernuni-hagen.de
m.kaufmann@hslu.ch

Keywords

Big Data, CRISP4BigData, Cross Industry Standard Process for Big Data, IVIS4BigData, Reference Model for Big Data Management, BDM ^{cube}, Virtual Research Environments, CRISP-DM, Cross Industry Standard Process for Data Mining

Abstract

This paper offers the Cross Industry Standard Process for Big Data (CRISP4BigData) based on the Big Data Management Reference Model (BDM ^{cube}), the IVIS4BigData Reference Model and the Cross Industry Standard Process for Data Mining (CRISP-DM) to manage a Big Data Analysis Process and the needed resources (e.g. information resources, hardware and software resources, algorithms, libraries, experts). CRISP4BigData could be used as Projects Reference Model as well as Process Reference Model to process Big Data Projects and its analysis.

1 Introduction

The volume of data in enterprise data centers increases between 35 and 50 percent each year (Beath et al. 2012). Through new wide-ranging developments in the area of information- and sensor technology, beside increasing computational processing capabilities, especially in terms of calculation, intensive and analytics methods will be fostered. The collected mass data stem from the internal and external corporate division and consist of unstructured data such like text data, processing information, (database-) tables, graphics, videos (Beath et al. 2012), e-mails, feeds and sensor data (Freiknecht 2014).

The acquisition of decisional relevant mass data is known as “Big Data”. At the beginning, the term was defined by companies which had to handle an amount of fast growing data. The information technology research and advisory company Gartner Inc. defines the term as follows: “Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making” (Gartner 2015). In contrast the market research company IDC defines Big Data as follows: “Big Data technologies as a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis” (Carter 2011).

The analysis of Big Data is a process which could be managed by analysis process models like CRISP-DM (Cross Industry Standard Process for Data Mining), SEMMA (Sample, Explore, Modify, Model, and Assess), or the KDD (Knowledge Discovery in Databases). Due to the new requirements (e.g. Meta data enrichment, archiving, automatic expert finder) on analysis, following,

the new analysis process model called Cross Industry Standard Process for Big Data (abbreviated CRISP4BigData) will be offered.

2 State of the Art

2.1 Big Data Management Model (BDM ^{cube})

Kaufmann describes the Management of Big Data, in his paper Towards a Reference Model for Big Data Management (cf. Figure 1), as act of value creation closely linked to the management of business intelligence and "the optimization of these five aspects of data integration, data analytics, data interaction, and data effectuation as well as the successful management of the emergent knowledge in this process, which can be called data intelligence." (Kaufmann 2016)

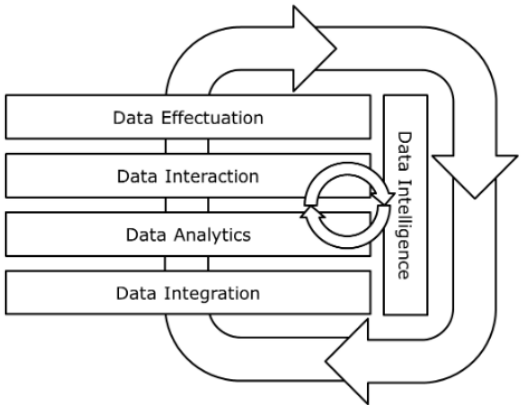


Figure 1. A knowledge-based Big Data Management Meta-Model (BDM ^{cube}) (Kaufmann 2016)

The offered BDM ^{cube}, shifts from an epistemic view of a cognitive system to a management view in a layer-based reference model. Kaufmann describes that the offered model "can be seen as a meta-model, where specific BDM models for a company or research process represent specific instances implementing certain aspects of the five layers."(Kaufmann 2016). The BDM ^{cube} can be used for the classification and enhancement of existing BDM Models, and it could be an inspiration to create or derive new BDM models for Big Data projects.

As already said, the BDM ^{cube} is based on five layers, used to serve as a frame of reference for the implementation, operation and optimization of BDM (Kaufmann 2016). Each of the five layers is described in detail:

- **Data Integration**: Defines the collection and combination of data from different sources into a single platform. The layer handles the involved database systems and the interfaces to all data sources with special care to the system scalability and the common characteristics of big data like volume, velocity, and the variety of data (sources).
- **Data Analytics**: Describes the transformation of raw data into information by the involvement of analytical processes and tools. Therefor big data, analytical and machine learning systems have to operate on a scalable, parallel computing architecture.
- **Data Interaction**: Is a layer which deals with analytical requirements of users and the results of the data analysis to create new organizational knowledge. Furthermore Kaufmann describes that "It is important to note that data analysis results are in fact nothing but more data unless users interact with them." (Kaufmann 2016)

- **Data Effectuation**: Describes the usage of data analytics results to create added values in products, services, and operations of organization.
- **Data Intelligence**: Defines the task of knowledge management and knowledge engineering over the whole data lifecycle, to deal with the ability of the organization, to acquire new knowledge and skills. Further the layer offers a cross-functional knowledge-driven approach to operate with the knowledge assets which are deployed, distributed, and utilized over all layers of Big Data Management.

2.2 IVIS4BigData Reference Model

Bornschlegl's extended IVIS4BigData Reference Model (cf. Figure 2 - IVIS4BigData Reference Model) is an enhancement and extension of the IVIS Reference Model in combination of the BDM^{cube} (Big Data Management Reference Model) (Kaufmann 2016) to support new conditions and opportunities with advanced visual interfaces for perceiving, managing, and interpreting Big Data analysis results to create new insights, emerge knowledge generation and improve the decision making.

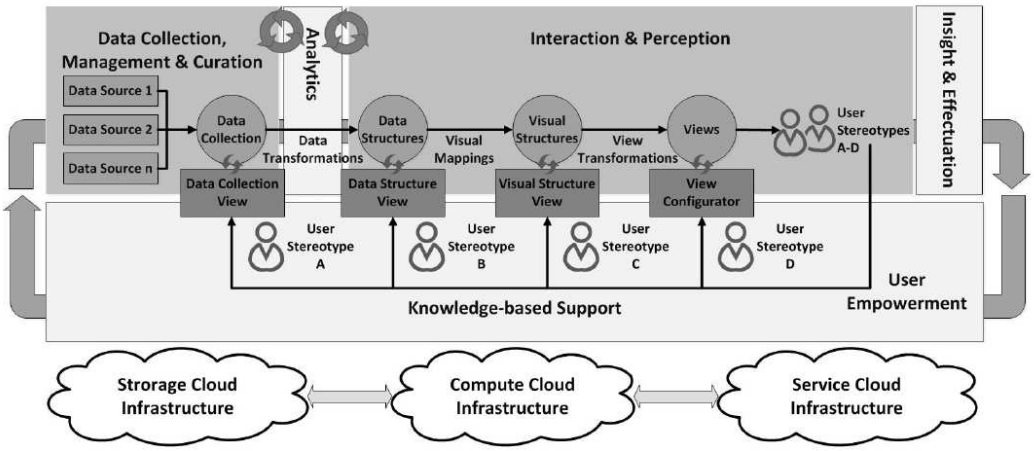


Figure 2. IVIS4BigData Reference Model (Bornschlegl 2016)

The IVIS4BigData Reference Model integrates the underlying BDM ^{cube} which illustrates different phases of Big Data Management. The IVIS Reference Model describes, as part of the BDM ^{cube}, the interactive part of the BDM life cycle. This means, the offered Reference Model supports an approach to deal with the connecting, collecting, processing and visualization with (raw) data from multiple interdisciplinary cross-domain and cross-organizational sources.

The model is divided into the Data **Collection, Management & Curation** layer, which connects, integrates, and manages data. The **Analytics** layer transforms the data from the sources into Data Structures which are represented within the **Interaction & Perception** layer. Furthermore the Data Structures will be transformed over Visual Mappings step into Visual Structures and View Transformations to create Views of Visual Structures (specified graphical parameters). **The Insight & Effectuation** layer creates and manages new insights from (existing) results of analysis and knowledge.

The major enhancement is located within the **Knowledge-based Support** and manages the different information requirements of various user stereotypes to offer "*the 'right' information, at the 'right' time, in the 'right' place, in the 'right' way to the 'right' person*" (Fischer 2012). Finally the

IVIS4BigData Reference Model offers a circulation functionality to use results as input for new process iterations.

2.3 Cross Industry Standard Process for Data Mining

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a process model developed in 1996 by the CRISP-DM consortium, existing of DaimlerChrysler (later Daimler-Benz), SPSS (later ISL), NCR Systems Engineering Copenhagen und OHRA Verzekeringen en Bank Groep B.V., with the target to handle the complexity of Data-Mining projects (Chapman 2000).

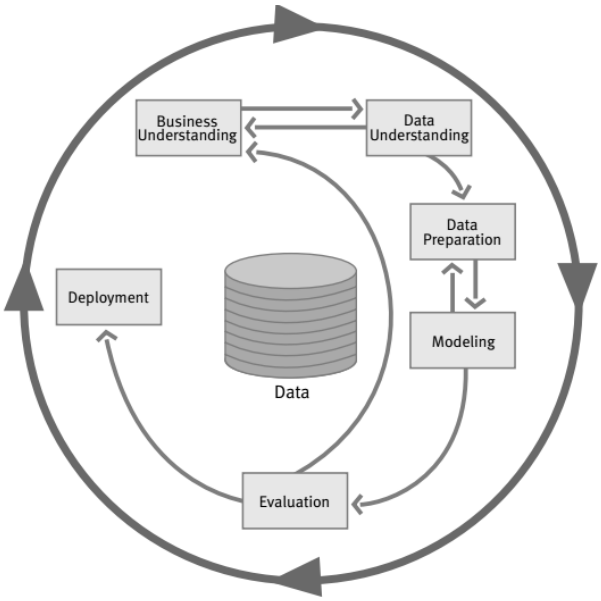


Figure 3. Cross Industry Standard Process for Data Mining

As in figure 3 shown the CRISP-DM consists of six, partly affiliated with each other and iterative phases. The base of the whole process are the phases **Business Understanding**, **Data Understanding**, **Data Preparation**, **Data Modeling**, **Evaluation** and **Deployment**, which are steering the project requirements and targets, as well as the Modelling, Evaluation and Deployment of a Data Mining Process (Chapman 2000).

The **Business Understanding** phase manages the projects objects and the requirements from the business point of view and designs a solution process (incl. projects plan) for the whole project (Chapman 2000).

Data Understanding starts with the collection of data and first ad-hoc analysis to become familiar with the data to identify problems (e.g. weak Data Quality) and to create first insights into the data (e.g. hidden information) (Chapman 2000).

The **Data Preparation** phase manages all needed activities to create a final dataset based on Raw Data collected in the Data Understanding phase. Data Preparation tasks could be performed at multiple times and various sequences to load, transform or clean data (Chapman 2000).

Modeling describes the application of various Modeling and Data Mining techniques to create new insights based on mathematical and statistical models. Sometimes it is necessary to adapt your initial dataset (based on the Data Preparation phase) to applicate other Modeling or Data Mining techniques (Chapman 2000).

The **Evaluation** phase reviews the built model to ensure a high quality of the analysis. Concerning this, the output of the analysis will be compared to the project and business objectives set at the Business Understanding phase (Chapman 2000).

The **Deployment** phase is the "go live" of the model. But it is necessary to organize and present this model in a way, customer or user, could use the output of the model to support the decision making process (e.g. Dashboards, Reports) (Chapman 2000).

3 Conceptual Modelling of Cross Industry Standard Process for Big Data Reference Model

CRISP4BigData Reference Model is based on Kaufmann's reference model for **Big Data Management** (Kaufmann 2016), Bornschlegl's **IVIS4BigData** reference model (Bornschlegl et al. 2016). IVIS4BigData was evaluated during an AVI 2016 Workshop in Bari. The overall scope and goal of the workshop was to achieve a road map, which can support the acceleration in research, education and training activities by means of transforming, enriching, and deploying advanced visual user interfaces for managing and using eScience infrastructures in VREs. In this way, the research, education and training road map will pave the way towards establishing an infrastructure for a visual user interface tool suite supporting VRE platforms that can host Big Data analysis and corresponding research activities sharing distributed research resources (i.e. data, tools, and services) by adopting common existing open standards for access, analysis and visualization. Thereby, this research helps realizing an ubiquitous collaborative workspace for researchers which is able to facilitate the research process and its Big Data Analysis applications (Bornschlegl et al. 2016).

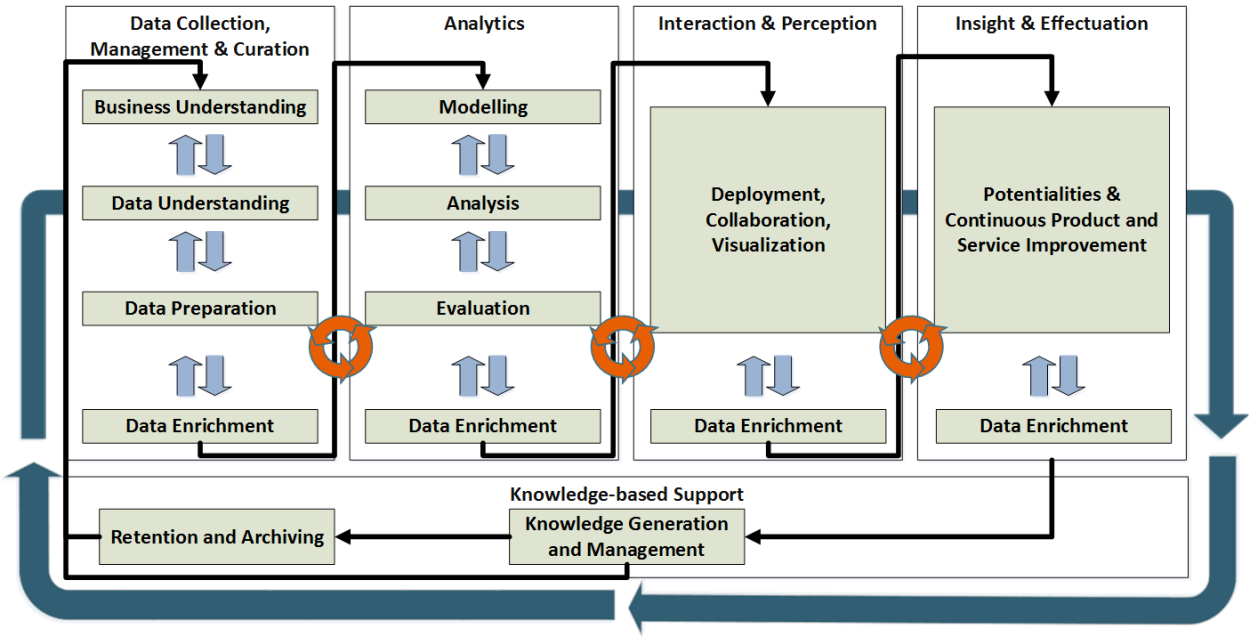


Figure 4. CRISP4BigData Reference Model Version 1.0

CRISP4BigData is an enhancement of the classical Cross Industry Standard Process for Data Mining (CRISP-DM) developed by the CRISP-DM consortium, existing of DaimlerChrysler (later Daimler-Benz), SPSS (later ISL), NCR Systems Engineering Copenhagen und OHRA Verzekeringen en Bank Groep B.V., with the target to handle the complexity of Data-Mining projects (Chapman et al. 2000).

The CRISP4BigData Reference Model (cf. Figure 4: CRISP4BigData Reference Model Version 1.0) is based on Kaufmann`s 5 phases of Big Data Management such as “Data Collection, Management & Curation”, “Analytics”, “Interaction & Perception”, “Insight & Effectuation”, and “Knowledge-based Support” and the standard four layer methodology of the CRISP-DM model. The CRISP4BigData Methodology (based on CRISP-DM Methodology) describes the hierarchical process model, consisting of a set of tasks disposed to the four layers **Phase**, **Generic Task**, **Specialized Task**, and **Process Instance**.

Within each of these sections are a number of phases (cf. Figure 5: CRISP4BigData Methodology based on CRISP-DM Methodology) (e.g. Business Understanding, Data Understanding, Data Preparation) analog to the standard description of the original CRISP-DM model enhanced by some new phases (e.g. Data Enrichment, Retention and Archiving).

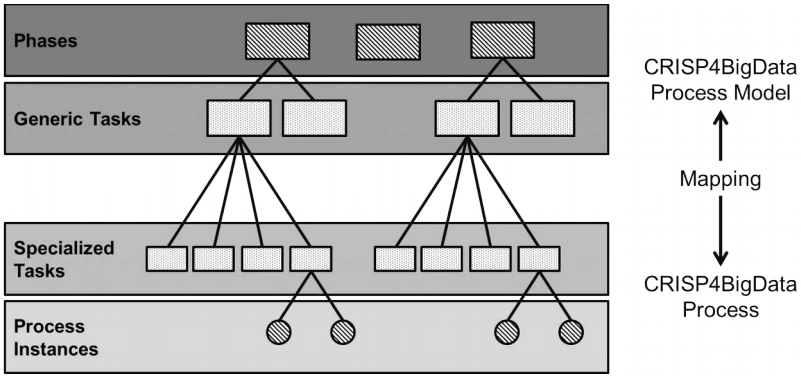


Figure 5. CRISP4BigData Methodology based on CRISP-DM Methodology

Each of these phases consists of second-level generic tasks. The layer is called generic because it is deliberated to be general enough to cover all conceivable use cases. The third layer, the Specialized Task, describes how actions, located in the generic tasks, should be processed and how they should differ in different situations. That means, the Specialized Task layer handles the way e.g. how data should be processed, “*cleaning numeric values versus cleaning categorical values, or whether the problem type is clustering or predictive modeling.*” (Chapman et al. 2000)

The fourth layer, the Process Instance, is a record set of running actions, decisions and results of an actual Big Data analysis. The layer describes that “*A process instance is organized according to the tasks defined at the higher levels, but represents what actually happened in a particular engagement, rather than what happens in general.*” (Chapman et al. 2000)

Following, the different Phases of the CRISP4BigData Reference Model are described in detail:

Business Understanding describes the goals and problem statements, derived from a project plan to develop a targeted deployment of Big Data Analysis Methods. The Data Understanding phase deals with the collection of internal and external data, the description of data types and source systems. The Data Preparation manages the preparation of the Data over a Transformation or Data Cleansing (or update, enlarge, and reduce) to increase the data quality. All Data Enrichment process elements in the CRISP4BigData have the task to enrich the data base with useful Meta Data according to the whole process, to get insight of the process information to increase the quality of the process or to update the Knowledge-based Support phase with information such as: Who uses which data or analysis methods? Who is a knowledge carrier? Where is the data from? Where is the data used?

The Modelling element describes the development and implementation of a statistical or mathematical model based on an adequate data model. The Analysis element is based on the statistical or mathematical model implemented in the Modelling element and describes the deployment of a Big Data analysis method or algorithms to analyze the available data (model). The Evolution rates the result of the analysis and the process. The phase evaluates also the precision, usefulness, novelty, and significance of the result. The Deployment, Collaboration, and Visualization phase deals with the deployment, visualization of the analysis result and manages the distribution of the right result to the right persons (collaboration).

The Potentialities & Continuous Product and Service Improvement element steers the management of potentialities and cares about a continuous improvement process to work out some new opportunities for the company. The process element Knowledge Generation and Management achieve that the generated insights and knowledge do not get lost. The Retention and Archiving step manages a long-term retention and archiving of the data, it manages also the classification of hot, cold and warm data (tiers). Finally, the circulations around the whole phases show that the CRISP4BigData reference model is not an on-time process. It is partially useful or needed to repeat the whole process to deal with new information, obtained during the project or to improve the whole process.

4 Conceptual VRE based System Architecture Supporting a Big Data Proof-of-Concept Implementation

To evaluate the offered CRISP4BigData Reference Model, supporting scientists and research facilities to maintain and manage their research resources, will be done by an implementation of a proof-of-concept implementation. The Implementation is based on the conceptual architecture for a VRE infrastructure called Virtual Environment for Research Interdisciplinary Exchange (cf. Figure 6: Conceptual Architecture for a VRE Infrastructure) mentioned by Bornschlegl´s paper "IVIS4BigData: A Reference Model for Advanced Visual Interfaces Supporting Big Data Analysis in Virtual Research Environments". (Bornschlegl forthcoming)

The CRISP4BigData Reference Modell will be implemented to handle the whole analytical process which is able to manage (automatically) all the needed information, analytical models, algorithms, library's knowledge resources, and infrastructure resources to get new insights and optimize products and services.

The offered VRE model is "*building on the concept of semantic information integration and mediation and corresponding mediator architectures to support information integration across borders of scientific knowledge domains*." (Bornschlegl forthcoming) The various scientific resources and disciplines are "*mediated by means of semantic integration of domain models on the level of the mediator and by means of domain adaptation on the level of the corresponding wrappers.*" (Bornschlegl forthcoming)

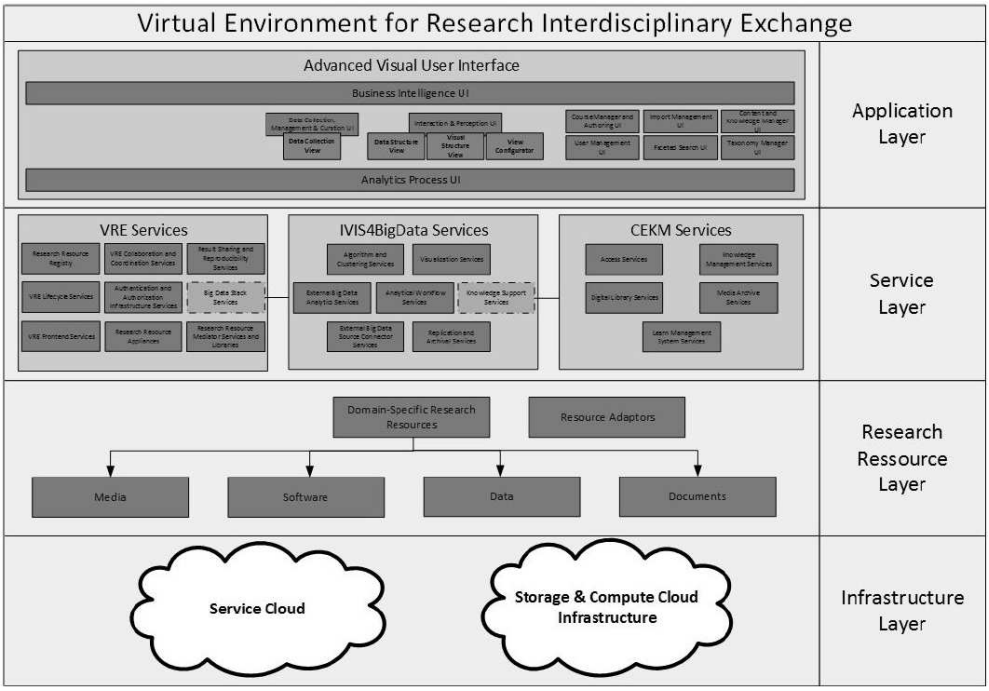


Figure 6. Conceptual Architecture for a VRE Infrastructure (Bornschlegl forthcoming)

The Infrastructure architecture is built on the concept of **Infrastructure as a Service (IaaS)** and uses the EGI Federal Cloud Infrastructure to use the positive benefits of it (e.g. scalability, elasticity, self-service, and accounting capabilities as defined by NIST). (National Institute of Standards and Technology 2011)

5 Related Work

5.1 gCube

"Data sharing has been an emerging topic since the 1980's. Science evolution – e.g. data-intensive, open science, science 2.0 – is revamping this discussion and calling for data infrastructures capable of properly managing data sharing and promoting extensive reuse." (Candela 2015) Therefor the gCube (Figure 7 - gCube System Architecture (Candela 2015) software system was designed to create and operate an innovative typology of data infrastructure. The development of gCube was benefited through "Grid, Cloud, digital library and service-orientation principles and approaches …" to deliver a sustainable and holistically data management infrastructure (as-a-service) (Candela 2015).

In detail, gCube is a core technology supporting e-infrastructures (e.g. SURFNet (information), GriPhyN (services), EGEE (computing and storage)) as service-oriented application framework, supporting scientists to deploy declaratively and dynamically Virtual Research Environments.

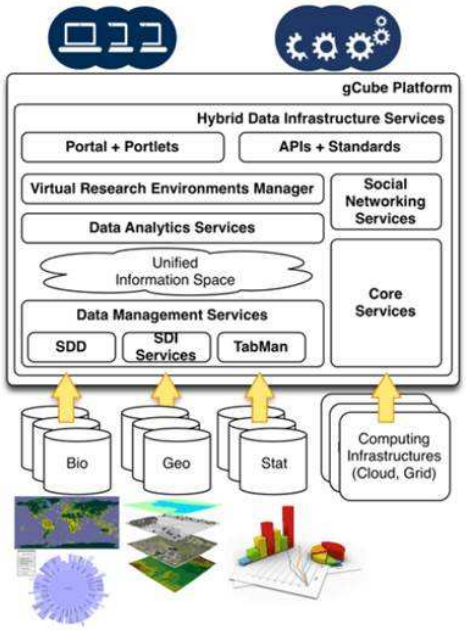


Figure 7. gCube System Architecture (Candela 2015)

gCube provides users with services for access and aggregates on-demand content resources, application services and computing resources (Candela 2008). The System "*also monitors the shared resources during the lifetime of the VRE, guaranteeing their optimal allocation and exploitation. Finally, it provides mechanisms to easily create dedicated Web portals through which scientists can access their content and services.*" (Candela 2008)

In a future peculiarity, gCube provides access to species data, geospatial data, statistical data and semi-structured data. The Data is located by various data providers and information systems and could be processed via web-based graphical user interfaces and protocols (e.g. OAI-PMH, SDMX) (Candela 2015).

6 Conclusion and future work

The evaluation of the offered model will be done by defined use cases part of the EU co-funded project SenseCare (SenseCare 2016). The specific use case is based on an unstructured data set of XML-based medical documents (circa 40 gigabyte) and is going to be transferred in a suitable ontology.

Therefor the process model is going to be implemented within a cluster setup, based on virtual machines located in the EGI Federated Cloud infrastructure and on local virtual machines located at servers at the University of Hagen. The cluster is based on several Apache and Apache Hadoop components.

The CRISP4BigData will be additionally implemented with an Apache based workflow-engine to automate and describe the whole analysis process. The description of the analysis process based on CRISP4BigData is going to be implemented with a graphical user interface.

7 Acknowledgment & Disclaimer

This publication has been produced in the context of the SenseCare project. The project has received funding from the European Union's Horizon 2020 research and innovation

programme under grant agreement No 690862. 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.

8 References

Beath, C., Becerra-Fernandez, I., Ross, J., Short, J. (2012): Finding Value in the Information Explosion. In MITSloan Management Review. Available online at http://sloanreview.mit.edu/article/finding-value-in-the-information-explosion/, checked on 5/25/2015.

Bornschlegl, M. X., Berwind, K., Kaufmann, M., Hemmje M. (2016): Towards a Reference Model for Advanced Visual Interfaces Supporting Big Data Analysis in Virtual Research Environments, Proceedings on the International Conference on Internet Computing (ICOMP): 78-81. Athens: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Bornschlegl, M. X., Berwind, K., Kaufmann, M., Engel, F. C., Walsh, P., Hemmje M., Riestra, R. (forthcoming): IVIS4BigData: A Reference Model for Advanced Visual Interfaces Supporting Big Data Analysis in Virtual Research Environments

Bornschlegl, M. X., Manieri, A., Walsh, P., <atarci, T., Hemmje. L. H. (2016). Road Mapping Infrastructures for Advanced Visual Interfaces Supporting Big Data Applications in Virtual Research Environments. In Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '16), Paolo Buono, Rosa Lanzilotti, and Maristella Matera (Eds.). ACM, New York, NY, USA, 363-367. DOI=http://dx.doi.org/10.1145/2909132.2927471

Candela, L., Castelli, D., Pagano, P. (2008). gCube: A Service-Oriented Application Framework on the Grid. http://ercim-news.ercim.eu/en72/rd/gcube-a-service-oriented-application-framework-on-the-grid. checked on: 10/28/2016

Candela, L., Pagano, P. (2015). Cross-disciplinary Data Sharing and Reuse via gCube. http://ercim-news.ercim.eu/en100/special/cross-disciplinary-data-sharing-and-reuse-via-gcube. checked on: 10/28/2016

Carter, P. (2011): Future Architectures, Skills and Roadmaps for the CIO. Edited by IDC. IDC. Available online at http://www.sas.com/resources/asset/BigDataAnalytics-FutureArchitectures-Skills-RoadmapsfortheCIO.pdf, updated on 2011, checked on 7/1/2016.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C.; Wirth, R. (2000): CRISP-DM 1.0. Step-by-step data mining guide. Edited by CRISP-DM Consortium. Available online at ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf, checked on 6/9/2016.

Fischer, G (2012). Context-aware systems: The 'right' information, at the 'right' time, in the 'right' place, in the 'right' way, to the 'right' person. In Proceedings of the International Working Conference on Advanced Visual Interfaces, ser. AVI '12. New York, NY, USA: ACM, pp. 287–294.

Freiknecht, J. (2014): Big Data in der Praxis. Lösungen mit Hadoop, HBase und Hive; Daten speichern, aufbereiten, visualisieren. München: Hanser.

Gartner, Inc. (2015): Gartner - IT-Glassary Big Data. Available online at http://www.gartner.com/it-glossary/big-data/, checked on 8/1/2015.

Kaufmann, M. (2016). Towards a reference model for big data management. Research report, Faculty of Mathematics and Computer Science, University of Hagen, retrieved July 15, 2016 from: https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00000583

National Institute of Standards and Technology. The nist definition of cloud computing. Recommendations of the National Institute of Standards and Technology. http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf, checked on 10/28/2016

SenseCare (2016). SenseCare - Sensor Enabled Affective Computing for Enhancing Medical Care, checked on 08/30/2016.

Social Network Support for an Applied Gaming Knowledge Management Ecosystem

Munir Salman [1,4], Jana Becker [2], Michael Fuchs [3], Dominic Heutelbeck [2], Matthias Hemmje [1]

[1] FernUniversität in Hagen, Faculty for Multimedia and Computer Science, Hagen, Germany

[2] Forschungsinstitut für Telekommunikation und Kooperation e.V. (FTK), Dortmund, Germany

[3] Wilhelm Büchner University of Applied Science, Pfungstadt, Germany

[4] Munich Center for the Economics of Aging (MEA)
at the Max Planck Institute for Social Law and Social Policy

Munir.Salman@fernuni-hagen.de

Keywords

Social Networking, Knowledge Management and Transfer, Applied Gaming, Digital Ecosystem

Abstract

The European (EU)-based industry for non-leisure games (Applied Games, AGs or formally called serious games) is an emerging business. As such it is still fragmented and needs to achieve critical mass to compete globally. Nevertheless, its growth potential is widely recognized and even suggested to exceed the growth potential of the leisure games market. To become competitive the relevant applied gaming communities (applied game researchers, developers, customers, and players) require support by fostering the generation of innovation potential. The European project Realizing an Applied Gaming Ecosystem (RAGE) (Netvision, 2015) is aiming at supporting this challenge. RAGE will help to seize these opportunities by making available an interoperable set of advanced Applied Game (AG) technology assets, as well as proven practices of using asset-based AGs in various real-world contexts. As described in (Salman et al., 2015b), RAGE will finally provide a centralized access to a wide range of applied gaming software modules, relevant information, knowledge and community services, and related scientific documents, taxonomies, media, and educational resources within an online community portal called the RAGE Ecosystem. This Knowledge Management Ecosystem (KM-ES) is based on an integrational, user-centred approach of Knowledge Management (KM) and Innovation Processes in the shape of a service-based implementation (Becker et al., 2015). This will support information and knowledge sharing, as well as persistency of social interaction threads and Know-how transfer within Social Networking Sites (SNSs) and Groupware Systems (GWSs) that are connected to the RAGE Ecosystem. In this paper, we will review the integration of several Social Networking Platforms (SNPs) (e.g., LinkedIn and Twitter), as well as Groupware Systems (GWSs) such as the Scientific Publication and Presentation Platforms (SPPs) Mendeley (Netvision, 2016a) and SlideShare ("SlideShare.net," 2016), along with GitHub software repository ("Build software better, together," 2016) and the StackExchange network ("Hot Questions - Stack Exchange," 2016) into the RAGE KM-ES. The integration approach of SNPs, SPPs, and GWSs into the RAGE KM-ES based on KM and Innovation Processes as described in (Salman et al., 2016b), (Salman et al., 2015b) and (Salman et al., 2016a) enhance the RAGE KM-ES with Social Networking Features (SNFs) such as Share, Like, Review and Follow. On the other hand, this will allow for automating repetitive tasks, reducing errors, and speeding up time consuming tasks. Furthermore, this paper reviews related authentication, authorization, access, and information integration challenges in the context of the RESTful web service architecture (Benson and Grieve, 2016) and (Gusmeroli et al., 2013). Our evaluation which is based on relevant use cases and scenarios demonstrates that such integration of several SNPs and GWSs into the RAGE KM-ES is a feasible framework, on the one hand, to enhance KM and Know-how transfer in an Applied Gaming Ecosystem (AGE) and on the other hand to reduce the fragmentation among applied gaming communities.

1 Introduction and motivation

The EU-based industry for Applied Games (AGs) is an emerging business. As such it is still fragmented and needs to achieve critical mass to compete globally. Nevertheless, its growth potential is widely recognized and even suggested to exceed the growth potential of the leisure games market. Though, the launch of innovative products for SMEs of the AG industry constitutes an enormous challenge considering the global competition combined with limited budgets. They

need strategies to have the crucial competitive advantage of being faster than others (Haß, 1983; Paukert et al., 2011). Accelerating the discovery of new scientific findings, the technical realization and the market launch (Haß, 1983) is increasingly dependent on the use of advanced information and knowledge technology for building environments that support the innovation process systematically and efficiently (Specht et al., 2002). Such environments depend on a number of advanced Knowledge Management (KM) technologies and processes and have to adapt to a wide variety of innovative practices, cultures, organizational context and application areas, where innovation takes place. Independent of the domain, innovation is a knowledge-intensive process. (Paukert et al., 2011) The RAGE project (Netvision, 2015) is aiming at supporting this challenge. RAGE will help to seize these opportunities by making available an interoperable set of advanced technology assets, tuned to AG, as well as proven practices of using asset-based AGs in various real-world contexts. This will be achieved by enabling a centralized access to a wide range of AG software modules, information, knowledge and community services, as well as related documents, publications, media, and educational resources within the RAGE KM-ES. Furthermore, the RAGE project aims to boost the collaboration of diverse actors in the AG environment. The main objectives of the RAGE KM-ES are to allow its participants to get hold of advanced, usable gaming assets (technology push), to get access to the associated business cases (commercial opportunity), to create bonds with peers, suppliers, and customers (alliance formation), to advocate their expertise and demands (publicity), to develop and publish their own assets (trade), and to contribute to creating a joint agenda and road-map (harmonization and focus). Therefore, the RAGE project is a technology and know-how driven research and innovation project. Its main driver is to be able to equip industry players (e.g., game developers) with a set of AG technology resources (so-called Assets) and strategies (i.e., know-how) to strengthen their capacities to penetrate an almost new market (non-leisure) and to consolidate a competitive position. Figure 1 represents the positioning of the project in the spectrum from 'theory to application'.

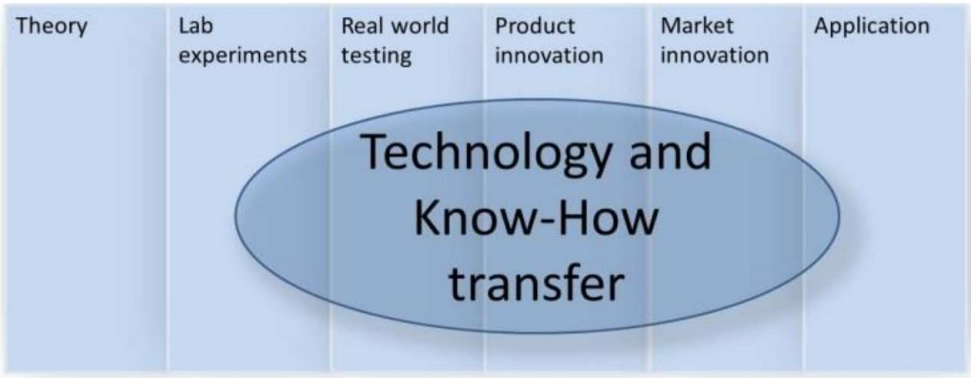


Figure 1. Technology and Know-How transfer (Netvision, 2015)

In consequence, the RAGE KM-ES and its integration with social networks of game- research-, game-developing-, gaming-, and AG communities will on the one hand become an enabler to harvest community knowledge and on the other hand it will support the access to the RAGE KM-ES as an information and knowledge resource for such communities. The AG sector as an upcoming business market is at present characterized by weak interconnectedness, limited knowledge exchange, absence of harmonising standards, limited specialisations, limited division of labour, and insufficient evidence of the products' efficacies (Sánchez et al., 2013; Stewart et al., 2013). The industry is scattered over a large number of small, diverse, independent players, niche products and of course specialists.

Because of limited collaborations of industries and limited interconnections between industry and research, applied gaming companies display insufficient innovation power and size to open up new markets (e.g. schools, business, governments) (Stewart et al., 2013). To support the development and growth of this branch the RAGE KM-ES will foster the merging of the heterogeneous AG communities by providing an effective knowledge and innovation management service tool. As a single entry point for AG, the RAGE KM-ES will realize centralized access to a wide range of AG software modules, services and resources by the arrangement of a well-managed and structured asset repository, digital library, and media archive system. The resulting material in the Ecosystem, particularly the textual resources, will be semantically annotated to support searching and access. Therefore, Social Network Analysis (SNA) by means of applying technologies for Natural Language Analysis (NLA) for discourse analysis will be used. Besides, the Ecosystem will arrange workshops and offer training courses on an online training portal, covering training for developers and educators in order to amplify applied gaming uptake. The aim will be to support the self-sustainable production of assets and documentation, training material, workshops and collaboration activities. In addition the social dimension of the RAGE KM- ES will be supported by community tools for collaboration, annotation, creativity and matchmaking (Salman et al., 2015a). Furthermore the Ecosystem will serve as an interactive knowledge and content management platform and provide a diverse set of services across the knowledge value chain (Salman et al., 2015a). In the remainder of this paper, section 2 describes related work to similar Ecosystems based on (KM) and Innovation Processes. The section 3 is about state of the art in science and technology. Furthermore, Section 4, more specifically, reviews the implementation and integration possibility of Mendeley and SlideShare into the RAGE KM-ES using their Application Programming Interface (API). Finally, the paper will present conclusions and future works.

2 Related work

The RAGE KM-ES was built upon the Educational Portal (EP) technology and application solution, which was developed by the software company GLOBIT ("Globit," 2016) that already was used in APARSEN (Schrimpf, 2014). APARSEN (Alliance Permanent Access to the Records of Science in Europe Network) was an EU-funded project within the digital preservation area with the goal to create a virtual research centre in digital preservation in Europe. The so-called EP tool-suite offers a wide variety of tools and is currently extended by Research Institute for Telecommunication and Cooperation (FTK) within the RAGE project into an Ecosystem Portal (EP) tool suite (Binh Vu, 2015a). This includes a web based, user-friendly User and Community Management (UCM) including an advanced Contact and Role Management (CRM) based on MythCRM (Binh Vu, 2015b), as well as KM support in the form of Taxonomy Management (TM) support and semi-automatic taxonomy-based Content Classification (CC) support (Nawroth et al., 2015),(Swoboda, 2014), as well as a Learning Management System (LMS) based on Moodle ("Lernerfolg mit Moodle," 2016). The EP is based on Typo3 ("TYPO3," 2016) and, therefore, can be extended with the help of Typo3 extensions. Our work will establish the new EP module Community & Social Network Support (CSNS) on the basis of a so-called Agile Application Programming Interface (AAPI), which facilitates the connectivity to a wide range of SNPs and GWSs. Furthermore the conceptual approach and benefits of Communities of Practice (CoP) by Lave and Wenger (Lave and Wenger, 1991) and the extensions to online CoPs (OCoPs) ("Online Communities of Practice | IDOE," 2016) respectively virtual CoPs (VCoPs) (Dubé et al., 2005) and the linking to KM (Kimble and Hildreth, 2006) are taken into account. Defined by people, especially practitioners, in a

shared domain engaging in a process of collective learning (Wenger, 2011), the RAGE KM-ES is creating and supporting the digital environment for a VCoP in the domain AG to enable successful asset-based (serious) game development and commercialization. Successful examples are the Community Grids for Learning (CGfL) (Netvision, 2016b), the NSW CoP for ICT professionals (Netvision, 2016c) or the IBM CoPs (Gongla and Rizzuto, 2001). A well-known similar Ecosystem, that is not especially focused on the domain of AG, is GitHub ("Build software better, together," 2016). GitHub is a software development Ecosystem which currently has over 10 million users. Users can upload or start a software project and collaborate on its development with other users in the community (Dabbish et al., 2012). Currently active and past projects are retained and can be used as examples and sources of knowledge. GitHub demonstrates that an Ecosystem with a partial overlap of features to the RAGE KM-ES, but with a different domain, can be successfully adopted as collaboration environments. Another example for empirically supporting the potential of the RAGE KM-ES is the GALA project. The GALA project involves 31 European institutions and facilitates the cultivation and dissemination of academic applied games knowledge (Gloria and Roceanu, 2010). One of the relevant differences between RAGE and GALA is that RAGE aims to facilitate the business market rather than the academic community. However, both RAGE and GALA have overlap in residing in the domain of AG.

3 State of the Art in Science and Technology

Today, most SNPs and GWSs provide so-called Application Programming Interfaces (APIs) based on REST architecture (Fielding, 2000) for developers to integrate relevant SNFs of the SNPs and GWSs into their systems. Although, the SNPs and GWSs are different in their functionality, i.e., their social networking feature support, their software architecture for the communication with distributed other systems is similar. As shown in Table 1, most of the SNPs and GWSs offer RESTful APIs (Masse, 2011), which can be used for integration with other systems.

SNP and GWS	Support / API	Content	Authorization	Search / CRUD	Social featues
Mendeley	Yes	Publication	OAuth 2.0	Yes	Share, Fav.
SlideShare	Yes	Presentation	OAuth 2.0	Yes	?
GitHub	Yes	SW-Code	OAuth 2.0	Yes	-
StackExchange	Yes	Text (Q&A)	OAuth 2.0	Yes	Rate
LinkedIn	Yes	Documents	OAuth 2.0	Yes	Share
Twitter	Yes	Documents, text	OAuth 2.0	Yes	Share, Rate, Fav.
ResearchGate	No	-	-	-	-
Google Schoolar	No, but 3rd-party	Publication	?	-	-
XING	Yes	Documents	OAuth 1.0	Yes	?
Youtube	Yes	Video	OAuth 2.0	Yes	?

Table 1. Examples of SNPs and GWSs APIs

In the following, the description of the Mendeley RESTful API software architecture as described in ("Mendeley Developer Portal," 2016) , as well as the SlideShare RESTful API as described in ("SlideShare.net," 2016) will be cited as an exemplary, illustrative, and at the same time representative example. Table 2 shows some functions and features, which can be accomplished with the Mendeley and SlideShare RESTful APIs.

Some Mendeley APIs' functions / features ("Mendeley Developer Portal," 2016)	**Some SlideShare APIs' functions / features** ("SlideShare.net," 2016)
• Annotations (CRUD) • Document attributes (CRUD) • Catalog Search • Documents (CRUD) • Retrieving BiBTex documents • Documents Metadata (CRUD) • Folders (CRUD) • Groups (CRUD) • Profiles (CRUD)	• Get Slideshow Information • Get Slideshows By Tag • Get Slideshows By User • Slideshow Search • Get User Favorites • Get User Contacts • Get User Tags • Upload Slideshow • Favorite Slideshow

Table 2. Some Mendeley and SlideShare APIs' functions and features

SNPs and GWSs are becoming increasingly important (Xiang and Gretzel, 2010). This holds especially true for various SNFs like, e.g., sharing and posting new Social Media Content (SMC), rating, commenting, tagging, chatting and liking, following actors or celebrities, playing games etc. These SNFs are not only entertaining and exciting but also useful for learning and for information enrichment. Research has shown that distance education courses are often more successful when they develop CoPs (Barab and Duffy, 2000). Besides, Breslin et al. define in (Breslin and Decker, 2007) "The Social Semantic Web as a vision of a Web where all of the different collaborative systems and social network services, are connected together through the addition of semantics, allowing people to traverse across these different types of systems, reusing and porting their data between systems as required." The integration of heterogeneous data in different distributed systems based on the mediator/ wrapper approach, originally proposed by (Wiederhold, 1992), has been used in several projects, e.g. (Garcia-Molina et al., 1997; Heiler and Zdonik, 1990; Pitoura et al., 1995). Wiederhold described mediators as software modules that transparently encode domain-specific knowledge about data or subset of data and share abstraction of that data with higher layers of applications. Wrappers are software modules that translate data to a common data model. RAGE will use Semantic Web technologies as well as mediators/ wrapper software architecture in order to describe in an interoperable way users' profiles, social connections, and social media creation and sharing across different SNPs and GWSs, as well as within the RAGE KM-ES. Therefore, RAGE will be able to deliver well-grounded recommendation and mediation features to AG R&D communities.

4 Integration approach and implementation

The following section presents the main technical integration possibilities in the backend, as well as in the frontend. In this way, our integration approach and methodology is enabling us to differentiate between how to get access to resources and assets in the RAGE KM-ES from external SNPs and GWSs communities and how to push contents from the RAGE KM-ES to the external SNPs and GWSs in order to improve user acceptance of services provided by the RAGE KM-ES.

The description of the Mendeley and SlideShare API software architectures as described in ("Mendeley Developer Portal," 2016, "SlideShare Developer Portal," 2016) will be cited as an exemplary, illustrative, and at the same time representative examples. Figure 2 displays the concept of a bi-directional integration approach of the RAGE KM-ES with SNPs (e.g. LinkedIn and Twitter) and GWSs (e.g. Mendeley and SlideShare) using a REST API. Corresponding to this bi-directional integration approach, the Tight and Loose Coupling methodologies, as described in (Salman et al., 2015b), will be considered for achieving an integration of SNPs and GWSs into the RAGE KM-ES.

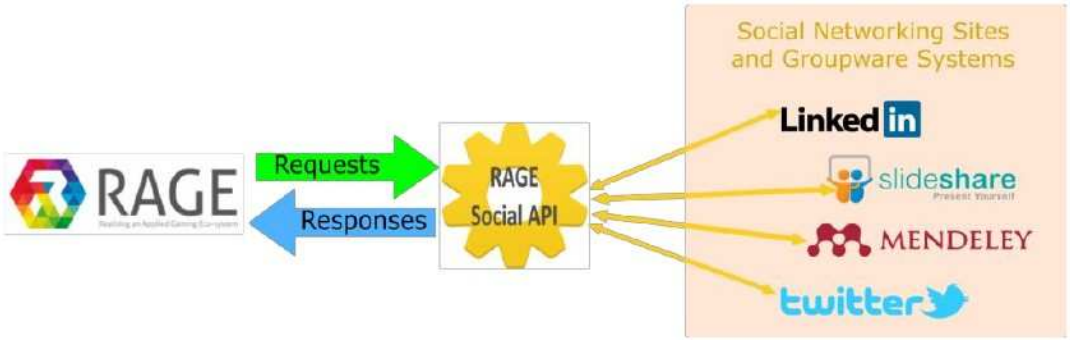


Figure 2. Integration Approach of the RAGE KM-ES with SNPs and GWSs

The Mendeley API is based upon the following standard:

- RESTful using the base URL: https://api.mendeley.com.
- JSON: The response data is delivered in JSON format.
- HTTPS: All requests must be send using TLS/SSL connections.
- OAuth 2.0 is used to authenticate and authorize all Mendeley API requests.
- CORS: Cross original resource sharing is enabled on all API requests. CORS is a mechanism to enable client-side cross-origin requests (Netvision, 2016d).

The following code illustrates an example for capturing data from the Mendeley System into the RAGE KM-ES using the Mendeley JavaScript Software Development Kit (SDK) (Netvision, 2016e). Each call will either resolve with some data or reject with the original request and the API response.

```
MendeleySDK.API.documents.list().done(function(docs) {
      console.log('Success!');
      console.log(docs);
}).fail(function(request, response) {
      console.log('Failed!');
      console.log('URL:', request.url);
      console.log('Status:', response.status);
});
```

The SlideShare API, similar to the Mendeley API, is based on the REST model, JSON, HTTPS, and OAuth 2.0. It supports the following actions (Netvision, 2016f):

- Upload, edit and delete slideshows.
- Retrieving slideshow information by user, tag, or group.
- Retrieving groups, tags, and contacts by user.
- Search slideshows.

The following code presents an example of uploading a presentation from the RAGE KM- ES into the SlideShare GWS for the logged in user.

```
public function UploadSlideShare($username, $password, $file, $title,
$description, $tags,
      $makeSrcPublic = false, $makeSlideShowPrivate = true,
            $generateSecretUrl = true,
                  $allowEmbeds = true, $shareWithContacts = false) {
                        $requestParams = array('username' => $username,
                              'password' => $password,
                              'slideshow_title' => $title,
                              'slideshow_srcfile' => $file,
                              'slideshow_description' => $description,
                              'slideshow_tags' => $tags,
                              'make_src_public' =>
                              ConvertBoolToSlideShareString(
                              $makeSrcPublic),
                              'make_slideshow_private' =>
                              ConvertBoolToSlideShareString(
                              $makeSlideShowPrivate),
                        ...
                        $request = 'https://www.slideshare.net/
                        api/2/upload_slideshow';
                        echo($request);
...
```

Corresponding to the code, Figure 3 displays the interface for uploading presentation in the SlideShare GWS through the RAGE KM-EP after successful login.

The integration of the Scientific Publication and Presentation Platforms (SPPs) Mendeley and SlideShare into the RAGE KM-ES facilitates the seamless integration of relevant Social

Networking Features (SNFs) and Groupware Features (GWFs), such as (rate, like, comment, share, post, etc.) as described in (Salman et al., 2015b).

Figure 3. Upload a presentation from the RAGE KM-EP into the SlideShare

5 Conclusion and future work

In summary, it is a big advantage to aim at supporting the integration of SNPs and GWSs (e.g. LinkedIn, Mendeley, SlideShare, etc.) including relevant SNFs and GWFs, as well as content capturing, sharing, management, and dissemination support through their RESTful API into the RAGE KM-ES. In this way, the RAGE KM-ES will provide AG communities (such as technology providers, game developers and educators, game industries and researchers), and therefore SNPs and GWSs communities, too, an opportunity to interact, share and re-use content including corresponding knowledge resources, as well as communicate and collaborate using the RAGE KM-EP as a back-end community content and KM portal in addition to their favorite SNPs and GWSs. This will on the one hand facilitates to provide a wide range of supporting services in the field of knowledge transfer and -creation, to overcome low market access and small market share of small and medium sized companies in the AG market, to create new effective technology based assets in order to build new ingenious learning games. On the other hand it focuses on identifying collaboration opportunities between individuals and among groups, to support matchmaking and collaboration between stakeholders, and to identify and provide support for innovation opportunities and creativity efforts. That allows communities to engage themselves in a VCoP (Allee, 2000), create their own assets and post them to the Ecosystem's repository without major effort and to benefit from achieving (business) results. Therefore knowledge becomes an economic asset itself and the process of knowledge creation becomes a value added service. The challenge will be to build up a sustainable, accepted and trusted environment, with somebody taking the lead and responsibility for the KM-ES (cf. e.g., (Dubé et al., 2005)). With the design and development of a comprehensive approach as pursued with the RAGE Ecosystem, ethical issues need to be taken into account. The integration of users' SN profiles from different SNPs and GWSs, as well as the use of features carrying out analyses on top of Ecosystem user data have ethical implications in terms of privacy and data protection and require appropriate information and consent in the terms and conditions of use, as well as compliance to national and international data protection regulations. In the preceding section an example shows uploading a presentation to the SlideShare GWS, it illustrate how user can select options between presentation privacy (public/private) and downloading file. In this way the system can also control data access by respecting personal settings

which data should be available to others or the public. This means that data protection and privacy is already taken into account when the system is being designed based on an ethics-by-design approach (Gotterbarn et al., 1997).

The RAGE KM-ES supports therefore the interconnectedness, the knowledge exchange and the harmonization of standards of the applied gaming branch. Coupled with suitable business models the RAGE KM-ES could help the AG industry to drive strategy, support problem solving, build capabilities and knowledge competencies, cross fertilize ideas and increase opportunities for innovation to assert themselves against big games companies.

In the future the RAGE KM-ES will support a RESTful API to manage all registered functions, which could be called by external sources such as SNPs, GWSs and Mobile Applications to get access to the assets of the RAGE KM-ES. Besides, it's necessary to connect the RAGE KM-ES to an appropriate data access control management software based on user roll management such as MythCRM (Binh Vu, 2015b) in order to manage access rights e.g. to the RAGE assets.

6 Acknowledgment & Disclaimer


This publication has been produced in the context of the RAGE project. The project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 644187. 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.

7 References

Allee, V., 2000. Knowledge networks and communities of practice. OD Pract. 32, 4-13.

Barab, S.A., Duffy, T., 2000. From practice fields to communities of practice. Theor. Found. Learn. Environ. 1, 25-55.

Becker, J., Lankveld, G. van, Steiner, C., Hemmje, M., 2015. Realizing an Applied Gaming Ecosystem: Towards

Supporting Service-based Innovation Knowledge Management and Transfer, in: Proceedings of the 4th International Conference on Games and Learning Alliance, GALA 2015. Presented at the GALA Conference, Rome, Italy.

Benson, T., Grieve, G., 2016. The FHIR RESTful API, in: Principles of Health Interoperability. Springer International Publishing, Cham, pp. 349-359.

Binh Vu, D., 2015a. Realizing an Applied Gaming Ecosystem - Extending an Education Portal Suite towards an Ecosystem Portal (Master Thesis). Technische Universität Darmstadt, Darmstadt, Germany.

Binh Vu, D., 2015b. Customer Relationship Management based on Identity Management for Scientific Associations (Bachelor Thesis). Technische Universität Darmstadt, Darmstadt, Germany.

Breslin, J., Decker, S., 2007. The Future of Social Networks on the Internet: The Need for Semantics. IEEE Internet Comput. 11, 86-90. doi:10.1109/MIC.2007.138

Build software better, together [WWW Document], 2016. . GitHub. URL https://github.com (accessed 7.26.16).

Dabbish, L., Stuart, C., Tsay, J., Herbsleb, J., 2012. Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12. ACM, New York, NY, USA, pp. 1277-1286. doi:10.1145/2145204.2145396

Dubé, L., Bourhis, A., Jacob, R., 2005. The impact of structuring characteristics on the launching of virtual communities of practice. J. Organ. Change Manag. 18, 145-166. doi:10.1108/09534810510589570

Fielding, R.T., 2000. Architectural styles and the design of network-based software architectures. University of California, Irvine.

Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., Vassalos, V., Widom, J., 1997. The TSIMMIS Approach to Mediation: Data Models and Languages. J. Intell. Inf. Syst. 8, 117-132. doi:10.1023/A:1008683107812

Globit [WWW Document], 2016. URL http://globit.com/ (accessed 10.20.16).

Gloria, A., Roceanu, I., 2010. Serious Games in the Life Long Learning environment. Games and Learning Alliance Network of Excellence, in: In Proceedings of the 5th International Conference on Virtual Learning ICVL. Presented at the The 5th International Conference on Virtual Learning ICVL.

Gongla, P., Rizzuto, C.R., 2001. Evolving communities of practice: IBM Global Services experience. IBM Syst. J. 40, 842-862.

Gotterbarn, D., Miller, K., Rogerson, S., 1997. Software engineering code of ethics. Commun. ACM 40, 110-118.

Grupp, H., 1997. Messung und Erklärung des technischen Wandels - Grundzüge einer empirischen Innovationsökonomik. Springer Verlag, Berlin Heidelberg.

Gusmeroli, S., Piccione, S., Rotondi, D., 2013. A capability-based security approach to manage access control in the Internet of Things. Math. Comput. Model., The Measurement of Undesirable Outputs: Models Development and Empirical Analyses and Advances in mobile, ubiquitous and cognitive computing 58, 1189-1205. doi:10.1016/j.mcm.2013.02.006

Haß, H.-J., 1983. Die Messung des technischen Fortschritts. München.

Heiler, S., Zdonik, S., 1990. Object views: Extending the vision, in: Sixth International Conference on Data Engineering, 1990. Proceedings. Presented at the Sixth International Conference on Data Engineering, 1990. Proceedings, pp. 86-93. doi:10.1109/ICDE.1990.113457

Hot Questions - Stack Exchange [WWW Document], 2016. URL http://stackexchange.com/ (accessed 7.26.16).

Kimble, C., Hildreth, P., 2006. The Limits of Communities of Practice, in: Encyclopedia of Communities of Practice in Information and Knowledge Management. Idea Group Reference, Hershey, PA, pp. 327-334.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press.

Lernerfolg mit Moodle [WWW Document], 2016. URL http://moodle.de/ (accessed 10.20.16).

Masse, M., 2011. REST API Design Rulebook. O'Reilly Media, Inc.

Mendeley Developer Portal [WWW Document], 2016. URL http://dev.mendeley.com/ (accessed 10.16.16).

Nawroth, C., Schmedding, M., Brocks, H., Kaufmann, M., Fuchs, M., Hemmje, M., 2015. Towards Cloud-Based Knowledge Capturing Based on Natural Language Processing, in: Procedia Computer Science, 1st International Conference on Cloud Forward: From Distributed to Complete Computing. pp. 206-216. doi:10.1016/j.procs.2015.09.236

Netvision, 2016a. Free reference manager and PDF organizer | Mendeley [WWW Document]. URL https://www.mendeley.com/ (accessed 7.28.16).

Netvision, 2016b. Community Grids for Learning (CGfL) [WWW Document]. URL http://homepages.shu.ac.uk/~edsjlc/ict/becta/information_sheets/commun2.pdf (accessed 6.16.16).

Netvision, 2016c. About ICT - Comprac [WWW Document]. URL http://www.comprac.nsw.gov.au/communities-of- practice/ict-professionals/about-ict (accessed 6.16.16).

Netvision, 2016d. Cross-Origin Resource Sharing [WWW Document]. URL https://www.w3.org/TR/cors/ (accessed 4.17.16).

Netvision, 2016e. Mendeley/mendeley-javascript-sdk [WWW Document]. GitHub. URL https://github.com/Mendeley/mendeley-javascript-sdk (accessed 11.23.15).

Netvision, 2016f. SlideShare‽» Entwickler & API [WWW Document]. URL http://de.slideshare.net/developers (accessed 4.17.16).

Netvision, 2015. RAGE 2016 [WWW Document]. RAGE. URL http://rageproject.eu/ (accessed 2.26.16).

Online Communities of Practice | IDOE [WWW Document], 2016. URL http://www.doe.in.gov/elearning/online- communities-practice (accessed 10.19.16).

Paukert, M., Niederée, C., Hemmje, M., 2011. Knowledge in Innovation Processes, in: Encyclopedia of Knowledge Management.

Pitoura, E., Bukhres, O., Elmagarmid, A., 1995. Object Orientation in Multidatabase Systems. ACM Comput Surv 27, 141-195. doi:10.1145/210376.210378

Salman, M., Becker, J., Fuchs, M., Heutelbeck, D., Hemmje, M., 2016a. Enhancing Knowledge Management and Transfer in an Applied Gaming Ecosystem, in: The 16th European Conference on Knowledge Management. Presented at the ECKM 2016.

Salman, M., Fuchs, M., Vu, B., Heutelbeck, D., Hemmje, M., Becker, J., Brocks, H., Research Institute for Telecommunication And Cooperation, 2016b. Integrating Scientific Publication into an Applied Gaming Ecosystem, in: Proceedings of the 17th European Conference on Knowledge Management, ECKM 2016. Presented at the ECKM Conference, Global Science & Technology Forum (GSTF), Northern Ireland, Belfast. doi:10.5176/2251-1679_CGAT16.9

Salman, M., Heutelbeck, D., Hemmje, M., 2015a. Towards Social Network Support for an Applied Gaming Ecosystem, in: Proceedings of the 9th European Conference on Games Based Learning ECGBL 2015. Presented at the The 9th European Conference on Games Based Learning ECGBL 2015, Steinkjer, Norway, pp. 721-728.

Salman, M., Star, K., Nussbaumer, A., Fuchs, M., Brocks, H., Vu, D.B., Heutelbeck, D., Hemmje, M., 2015b. Towards Social Media Platform Integration with an Applied Gaming Ecosystem. Presented at the SOTICS 2015,The Fifth International Conference on Social Media Technologies, Communication, and Informatics, pp. 14-21.

Sánchez, R.G., Hauge, J.B., Oliveira, M., Fiucci, G., Rudnianski, M., Hansen, P.K., Riedel, J., Padrón-Nápoles, C.L., Brown, D., 2013. Business Modelling and Implementation Report 2. GALA Network of Excellence.

Schrimpf, S., 2014. APARSEN - Alliance Permanent Access to the Records of Science in Europe Network. 52-53.

SlideShare Developer Portal [WWW Document], 2016. URL http://de.slideshare.net/developers (accessed 10.18.16).

SlideShare.net [WWW Document], 2016. . www.slideshare.net. URL http://www.slideshare.net (accessed 7.26.16). Specht, G., Beckmann, C., Amelingmeyer, J., 2002. F&E-Management. Schäffer-Poeschel-Verlag, Stuttgart.

Stewart, J., Misuraca, G., Jacobs, A., Grove, F.D., Willaert, K., Schurmans, D., All, A., Mariën, I., Looy, J.V., Bleumers, L., 2013. The Potential of Digital Games for Empowerment and Social Inclusion of Groups at Risk of Social and Economic Exclusion: Evidence and Opportunity for Policy. Joint Research Centre, European Commission.

Swoboda, T., 2014. Towards effectivity augmentation of automated scientific document classification by continuous feedback (Masterthesis). Fernuniversität Hagen, Hagen.

TYPO3 [WWW Document], 2016. . TYPO3 - Enterp. Open Source CMS. URL http://typo3.org/ (accessed 10.20.16).

Wenger, E., 2011. Communities of practice: A brief introduction.

Wiederhold, G., 1992. Mediators in the architecture of future information systems. Computer 25, 38-49. doi:10.1109/2.121508

Xiang, Z., Gretzel, U., 2010. Role of social media in online travel information search. Tour. Manag. 31, 179-188. doi:10.1016/j.tourman.2009.02.016

Improved Data Centre Network Management Through Software-Defined Networking

J. Sherwin
Department of Computing
CIT – Cork Institute of Technology, Ireland
e-mail: jonathan.sherwin@cit.ie

C. J. Sreenan
Department of Computer Science
UCC – University College Cork, Ireland
e-mail: cjs@cs.ucc.ie

Keywords: Software-Defined Networking, Data Centre Networks, Network Management

Research in the area of Computer Networking has seen a surge in activity in recent years with the advent of Software-Defined Networking (SDN), and with the OpenFlow (McKeown et al., 2008) protocol in particular. SDN is the application of programmatic control over the resources and functions of a network in order to make the network more dynamically configurable to match the requirements of users and applications. OpenFlow is an open, standard protocol that is commonly used to implement SDN between network switches and a centralised controller, allowing the controller to install flow-rules on those switches in order to implement network policy. Network switches then simply forward data packets according to the flow-rules that have been installed.

SDN offers significant potential benefit for all aspects of network management – e.g. configuration, monitoring, performance and utilisation, and security. SDN has been applied in campus, data centre, wireless, sensor and mobile networks. Our interest is in performance, monitoring and analysis of flow-rules in Data Centre Networks (DCNs). Particularly relevant work done by other researchers includes: Planck (Rasley et al., 2014), where DCN switches are monitored for congestion, and flows are re-routed to achieve fine-grained traffic engineering; a study of latency in flow-rule change requests made by the controller being applied within the network switches (Kuźniar et al., 2015); Procera (Hyojoon and Feamster, 2013), which can translate event-driven network policies into flow-rule configurations for switches; NetSight (Handigol et al., 2014), which records information about all packets that match a subset of flow-rules in a network; and OFRewind (Wundsam et al., 2011), where a proxy between controller and switches records and replays sequences of flow-rule management messages. For a more detailed survey of research and identification of open issues in SDN for DCN management, see our technical report (Sherwin and Sreenan, June 2016).

We plan to create a number of components of a flow-rule management system for a DCN. The first component will target reducing flow setup delay caused by the latency of hardware switches in processing flow-rule changes as described in (Kuźniar et al., 2015). It will do this by initially forwarding flows through pre-configured tunnels originating and terminating at software switches, then setting up the appropriate flow-rules on hardware switches to establish a non-tunnelled path, and verifying that path is operational before rerouting the flow to the non-tunnelled path. An emulated testbed environment has been set up as shown in Figure 1. The authors of (Kuźniar et al., 2015) have kindly provided a script to mimic the delays of specific models of hardware switch using a software switch, and we are currently writing additional scripts to verify that we can

reproduce their results in our emulated environment. Once verified, the testbed will be the environment in which we develop and test our flow-rule management components. We plan to extend the testbed by adding real hardware switches to further validate our work.

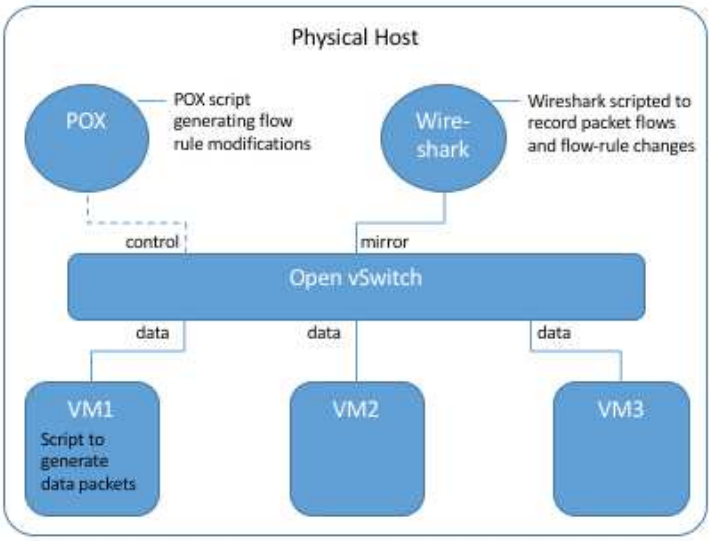


Figure 1: OpenFlow Test-bed Environment

DCNs can be comprised of thousands of switches carrying millions of flows. The flows often carry data belong to multiple tenants of the data centre. The data centre operator will have service level agreements with tenants that include minimum performance criteria. The network usually contains middle-boxes such as firewalls and load-balancers. Most data centres have a high level of server virtualisation, and an important feature facilitated by this is virtual machine (VM) migration. In order to be useful in a DCN, the components we develop will need to be scalable, low-latency, multi-tenant, middle-box friendly, and support VM mobility. To our knowledge, research efforts to date have not addressed all of these requirements for the functionality that our planned components will target.

Our research methodology is to develop prototype implementations targeting specific functionality, and to evaluate them in an emulated environment. This approach is commonly used for computer network research. For some components, we will be able to compare performance with other researchers' results for their attempts at providing similar functionality, for other components we will be providing functionality that has not been available up to now. For some of our planned components, we see opportunities to apply optimisation techniques for better performance, or less resource overheads. For all components, we will verify their usefulness in a DCN based on the criteria listed above.

In conclusion, our work is still at an early stage. We have reviewed SDN research from the last number of years and identified open issues in the area of DCN management. We have selected a few performance, monitoring and configuration issues on which to focus our own efforts. Having created a testbed environment, we are currently verifying it before we implement and evaluate components designed to address our selected issues.

References

Handigol, N., Heller, B., Jeyakumar, V., Mazi, D. and McKeown, N. (2014) 'I know what your packet did last hop: using packet histories to troubleshoot networks', Proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation, USENIX Association, pp. 71-85.

Hyojoon, K. and Feamster, N. (2013) 'Improving network management with software defined networking', Communications Magazine, IEEE, 51(2), pp. 114-119.

Kuźniar, M., Perešíni, P. and Kostić, D. (2015). What you need to know about SDN flow tables. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., Shenker, S. and Turner, J. (2008) 'OpenFlow: enabling innovation in campus networks', SIGCOMM Comput. Commun. Rev., 38(2), pp. 69-74.

Rasley, J., Stephens, B., Dixon, C., Rozner, E., Felter, W., Agarwal, K., Carter, J. and Fonseca, R. (2014) 'Planck: Millisecond-scale monitoring and control for commodity networks'. SIGCOMM 2014 - Proceedings of the 2014 ACM Conference on Special Interest Group on Data Communication, pp. 407-418.

Sherwin, J. and Sreenan, C. J. (2016) Software-Defined Networking for Data Center Managment - A Survey and Open Issues Technical Report UCC-CS-2016-24-06, Dept. of Computer Science, University College Cork, Ireland, June 2016.

Wundsam, A., Levin, D., Seetharaman, S. and Feldmann, A. 'OFRewind: enabling record and replay troubleshooting for networks', Proceedings of the 2011 USENIX Annual Technical Conference, USENIX Association, pp. 29-29.

Towards Synchronizing Heterogeneous Data Sources and Information Visualization

C. Danowski-Buhren [1], M. X. Bornschlegl [1], M. L. Hemmje [1], B. Schmidt [2]

[1] Department of Mathematics and Computer Science, University of Hagen, Germany

[2] Department of Geodesy, Bochum University of Applied Sciences, Germany

{christian.danowski, marco.bornschlegl, matthias.hemmje}@studium.fernuni-hagen.de
benno.schmidt@hs-bochum.de

Keywords

Bidirectional Visualization Pipeline, Information Visualization (IVIS), IVIS Reference Model, Mediator-Wrapper Pattern, Heterogeneous Data Sources, Synchronization

Abstract

This paper recommends an approach to synchronize 3D Information Visualization (IVIS) and heterogeneous data sources based on the Mediator-Wrapper architectural pattern. In addition, a bidirectional extension of the IVIS Reference Model, as proposed by Card et al. (1999), is introduced. Moreover, a prototypical implementation of the proposed IVIS infrastructure is presented utilizing the bidirectional WebSocket support of the open-source Spring framework. To establish an interactive 3D visual user interface, ISO standard X3D is embedded into an exemplary Web browser application with the help of the X3DOM JavaScript library.

1 Introduction

Due to technological progress and industrial as well as scientific research and development, the availability of data has increased over the past years. Through multi-purpose sensor systems or other data acquisition mechanisms within various discipline-specific domains (e.g. earth observation, weather/climate data), relevant application data is collected and stored within multiple heterogeneous data sources each day. The quantity and complexity of relevant *Big Data* requires new technologies with regard to data access, visualization, perception and interaction, as pointed out by Bornschlegl et al (2016). In addition, current eScience research resources (data, tools, ICT services) are often confined to computer science expert usage only. In consequence, it is a big challenge for research communities and industry to make those resources available for non-experts in an adequate way. Considering the users of data and ICT services within a Virtual Research Environment (VRE), their role shifts from passive consumers to more active participants. Keywords like *Web 2.0* (Campesato & Nilson, 2010) or the *Participative Web* (OECD, 2007) indicate that users intend to be empowered to actively contribute to application configuration and data management, including the possibility to create/manipulate content.

As an important means to improve the process of analysing and perceiving abstract data and to enable users to discover new insights, Information Visualization (IVIS) techniques can be employed. According to Spence (2014), exploration and perception of data can be performed in a cognitive efficient manner, if graphical representations of abstract data are offered through an intuitive direct-manipulative user interface. When using a three-dimensional depiction space, multiple properties of abstract data can be mapped to adequate visual attributes, such as position within an XYZ coordinate system, size or colour (Wiza, 2012). Through direct interaction with the displayed data, users may iteratively explore the data to discover new insights. As a widely

accepted conceptual model to transform abstract data to interactive visualizations, Card et al. (1999) developed the IVIS Reference Model, as depicted in Figure 1.

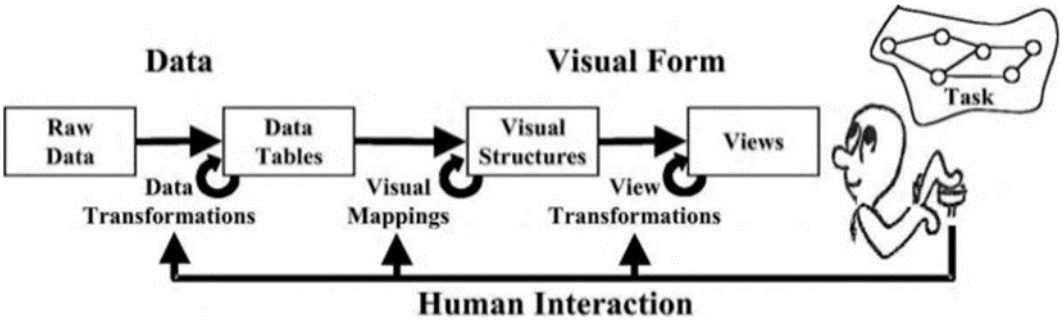


Figure 1. IVIS Reference Model designed by Card et al. (1999)

It comprises three dedicated user-configurable transformations from *Raw Data over Data-* and *Visual Structures* to final *Views* to enhance user cognition through visualizations. It emerged from research in the disciplines HCI, computer science, graphics, visual design, psychology and business methods to enable the discussion and comparison of IVIS systems (Card et al., 1999). However, with regard to the challenges and demands of current research, as motivated previously, the reference model reveals several deficits. First, it neglects the explicit consideration of multiple heterogeneous data sources as it only defines a single *Raw Data* component. Second, it focuses the unidirectional transformation from abstract data to views and was designed for static data. Hence, it misses means to reflect user-driven modifications of the underlying data at various processing stages of the pipeline, realizing a reverse transformation from view to data. Within the scope of rising User Empowerment and technological developments (e.g. cloud technologies, distributed computing, almost unlimited storage, computing performance), the reference model should be adapted to the present situation allowing dynamic data and bidirectionality.

To conquer challenges related to Big Data, Kaufmann (2016) proposes a reference model for Big Data Management (BDM), introducing several conceptual layers (Data Integration, -Analytics, -Interaction, -Effectuation and -Intelligence). A central interaction loop between the layers *Data Analytics* and *Data Interaction* reflects users iterative HCI to gain insight. By a global loop, Kaufmann includes the re-use of effectuated knowledge, as any achieved knowledge may serve as a new starting point for the next iteration of the whole model. Bornschlegl et al. (2016) build on Kaufmann's *BDM Reference Model* to augment the *IVIS Reference Model* from Card et al. (1999) with the layers of the BDM life cycle. The resulting **IVIS4BigData Reference Model** allows multiple heterogeneous data sources, cloud computing, user empowerment and user interaction and is depicted in Figure 2. Data interaction at the various stages of the embedded IVIS Reference Model is realized through dedicated views for a certain user stereotype. Their research aims to establish an information technology infrastructure for distributed Big Data Analysis in Virtual Research Environments (VREs) for use in eScience, Industrial Research, and Data Science Education to facilitate creative visual and cognitive efficient Human Computer Interaction (HCI). In addition, this paper presents an approach to conceptualize and implement a bidirectional IVIS infrastructure allowing users to modify the properties of data instances from within a 3D visual interface. The paper is outlined as follows: Section 2 concretizes the research goal and approach. A bidirectional extension of the IVIS Reference Model is proposed in section 3. The main part of the paper addresses the prototypical implementation of the proposed IVIS infrastructure, highlighting key architectural aspects in section 4. Section 5 demonstrates the usage of the infrastructure by

introducing an exemplar application scenario. Finally, a summary as well as hints for future work conclude the paper in section 6.

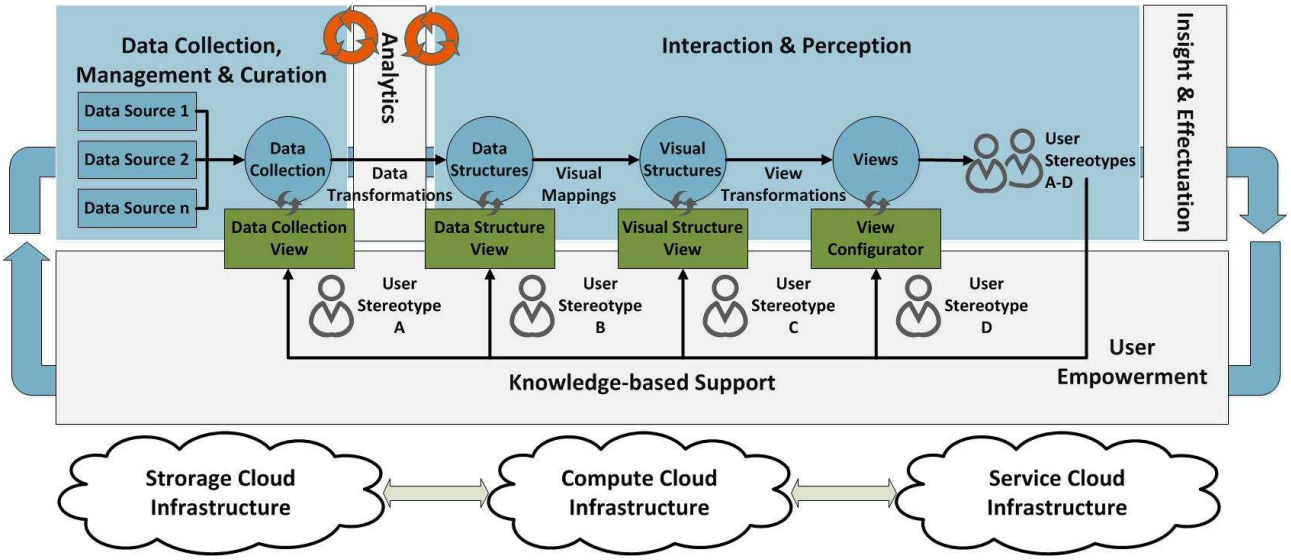


Figure 2. IVIS4BigData Reference Model (Bornschlegl et al., 2016)

2 Research Goal and Approach

The overall research goal is to develop a *universal, bidirectional 3D IVIS Web infrastructure* supporting Big Data in the application domains Information Visualization, Visual Analytics, Data Mining as well as Simulation and Analytics. The term *universal* indicates that the infrastructure is kept generic in order to adapt to different use case scenarios with distinct domain data and heterogeneous data sources. In addition, the visual mapping, which transforms abstract data to visual shapes, should be designed in a configurable way empowering users to easily switch between different visual representations. In relation to the IVIS4BigData Reference Model from Bornschlegl et al. (2016), this infrastructure may serve as bidirectional starting point enabling users to modify data instances through cognitive-efficient visual interfaces. Hence, the key aspect of the presented research approach is to design and implement a bidirectional IVIS infrastructure that allows the synchronization of Information Visualization and heterogeneous data sources. From a functional point of view, the following features have to be considered:

- **Server Side Information Visualization**: Allow clients to select abstract data and transform it into cognitive efficient visual representations.
- **Runtime Information Visualization**:
  - **Request Additional Data**: Users may request additional data from within an existing visualization. The retrieved data has to be transformed to visual representations and be integrated into the existing visualization.
  - **Modification of Data Instances**: Users may modify data properties through the visual interface. These modifications then have to be transmitted to the underlying data sources.
- **Synchronization**: As application content is dynamic, the properties of data instances might change over time. An occurring modification at data source level has to be broadcast to all connected clients in case they visualize the affected data instance. Without a

synchronization mechanism, the client might work on outdated data causing an inconsistent state between data source and visualization.

As initial research approach, suitable scientific approaches and contemporary (Web) technologies were identified, as elaborated by Danowski-Buhren et al. (2016). From a scientific point of view, in addition to the *IVIS Reference Model*, the *Mediator-Wrapper architectural pattern* should be employed, as it provides a homogeneous query/access interface for heterogeneous data sources (Wiederhold, 1992). A central *Mediator* component semantically integrates diverse data sources, each controlled by a dedicated *Wrapper* component, based on a unified global data schema and translates incoming queries against the global schema into appropriate subqueries against the local data schema of each individual wrapper / data source. With regard to modern (Web) technologies, Scene Description Language X3D/X3DOM and the WebSocket protocol have been identified. X3D is an ISO standard recommended by the Web3D Consortium (2016) to model interactive 3D multimedia content using XML syntax. The contents of a virtual scene are organized as a hierarchical structure called scene-graph. X3DOM (Behr et al., 2009) is based on X3D and integrates the nodes of the X3D scene graph into the DOM of a Web browser. In opposite to standalone X3D this approach has several advantages. First, as the scene-graph is embedded into the Web browser's DOM, it can be accessed natively using JavaScript. Second this allows rendering of the scene contents using WebGL, which is natively supported by modern Web browsers. This reduces the client-side need to install an additional Web browser plugin to view and interact with X3D content. The WebSocket is a full-duplex bidirectional Web protocol that in contrast to HTML establishes a permanent bidirectional communication channel in a client-server environment (Wang et al., 2013). It may serve as a key component of a bidirectional IVIS infrastructure, where both clients and server need to send messages to the other side at arbitrary points of time.

In summary, to develop a state-of-the-art 3D bidirectional IVIS infrastructure, the previously introduced key components represent applicable base technologies and approaches. Considering remaining challenges, the identified base technologies have to be coupled within the proposed infrastructure. In addition, to anchor the concept of a bidirectional visualization pipeline, the IVIS Reference Model from Card et al. (1999) should be extended to allow dynamic data and user-driven modification of data aspects explicitly. Consequently, this paper continues with the presentations of this extension.

3 Bidirectional Extension of IVIS Reference Model

The key aspect of the conceptual modelling is the bidirectional refinement of the IVIS Reference Model. In essence, abstract information must not only be transformed to visualizations as a unidirectional process. Instead, the reverse transformation from view to data has to be considered, allowing clients to persist modifications of the visualized data instances by updating the data sources themselves (North et al., 2010). Figure 3 illustrates the proposed extension. In addition to the stages of classical IVIS Reference Model, the component *Data Sources* has been added to support heterogeneous data sources. It is accompanied by a new transformation called *Data Generation and Selection* to reflect the dynamic nature of *Raw Data*. In particular, data might not only be selected but may be generated at runtime (e.g. by complex simulation or computation algorithms). Another major addition is the red backwards arrow, which represents user-triggered modifications of the various data aspects. Hereby, the focus of the presented research approach is to allow the modification of data instances (e.g. change properties of existing instances or create completely new instances).

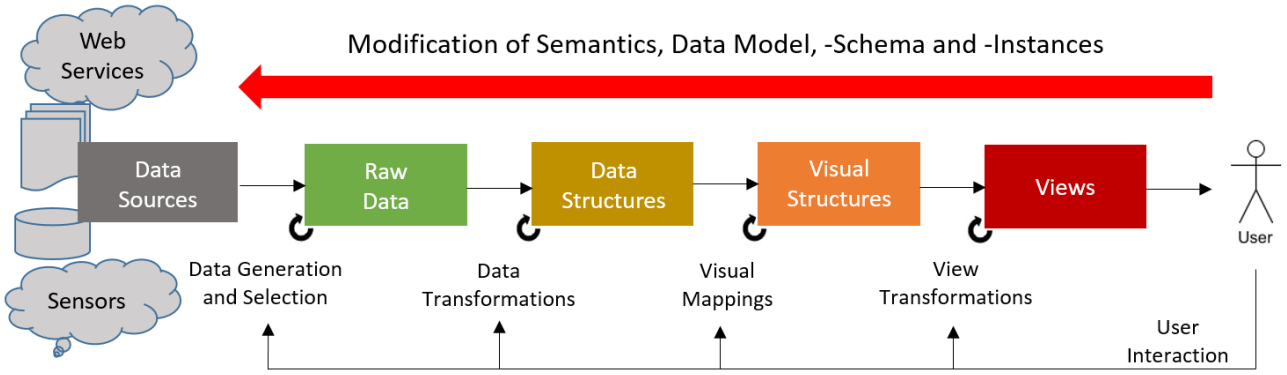


Figure 3. Extension of the IVIS Reference Model supporting dynamic data and user-driven modification of data aspects

In conclusion, the presented refinement of the IVIS Reference Model allows dynamic data and empowers users to modify it through the visual interface. Hence, this extension considers demands of current research topics in the fields of the "Participative Web" (OECD, 2007) as well as Big Data management and visualization (Bornschlegl et al., 2016). With regard to the overall research goal, the reverse transformation from view to data constitutes the key aspect of the bidirectional IVIS infrastructure.

4 Prototype Implementation of bidirectional IVIS Infrastructure

The concept of a bidirectional Web-based 3D IVIS infrastructure should be verified by a prototypical implementation. For this reason, the subsequent sections present key aspects of the developed prototype comprising used base technologies (section 4.1) and architectural considerations (section 4.2).

4.1 Base Technologies

The server-side implementation was developed using the programming language *Java* and project management tool *Maven*. Moreover, the prototype builds on the open source *Spring framework* (Pivotal Software, 2016), which offers a well-structured Model-View-Controller (MVC) infrastructure for Web applications. As a benefit, Spring enables enhanced WebSocket support compared to a standalone Java WebSocket implementation. As the WebSocket protocol is still young, outdated Web browser versions or hardware might reject it. For this reason, Spring integrates *SockJS* within its WebSocket support as an HTTP fallback mechanism. If a certain client rejects WebSocket, Spring automatically switches to HTTP streaming or polling techniques via SockJS to ensure a stable connection (Stoyanchev, 2012).

As recommended by various WebSocket programming guides, the *STOMP* protocol is layered on top of WebSocket. STOMP is an acronym for *Simple Text Oriented Messaging Protocol* and standardizes message exchange through simple text-based content. In addition, when using a so-called *STOMP broker*, the STOMP realizes a publish-subscribe mechanism, where clients can subscribe to dedicated channels/endpoints. If the broker receives a message directed against a certain channel, the broker automatically delivers that message to all subscribed clients. This mechanism even supports the possibility of sending messages to single users, if they subscribe to individual endpoints/channels. Hence, from the viewpoint of the implemented IVIS infrastructure, the communication with clients can be delegated to the STOMP broker. As a prerequisite, dedicated STOMP endpoints/channels have to be defined, to which clients have to subscribe when they

connect to the application. Again, the Spring framework offers the definition and use of a simple built-in STOMP broker, including the specification of concrete STOMP endpoints. In conclusion, building on the enhanced MVC-infrastructure of Spring, which enables enhanced WebSocket support with HTTP fallback and client communication through a STOMP broker, the prototype implementation is based on powerful open source base technologies.

With regard to the client application, standard Web technologies *HTML*, *JavaScript* and *CSS* may be facilitated. Due to the server-side technologies STOMP and SockJS, the corresponding counterpart JavaScript libraries *stomp.js* (Mesnil & Lindsay, 2015) and sockjs.js (SockJS, 2016) have to be employed on the client-side in order to handle communication with the server-side STOMP broker (e.g. to subscribe to STOMP endpoints/channels and receive messages). To create an interactive 3D visual interface within client applications, the JavaScript library *X3DOM* (Fraunhofer-Gesellschaft, 2016) is used. As it integrates X3D scene graphs into the DOM of the Web browser, client interaction with and manipulation of scene contents can be implemented through standard HTML/JavaScript events like *onclick*.

4.2 System Architecture

Composing the introduced base technologies from the prior section, the system architecture of the implemented IVIS infrastructure prototype is revealed according to Figure 4. As universality is one of the main implementation goals, the architecture is designed in a generic manner supporting a variety of use cases and application scenarios. On the client side, an HTML page, which offers the functionality to embed X3DOM scene nodes, serves as the central user interface. The application server consists of multiple components. A so-called *3D IVIS Mediator* component comprises three layers. Due to building on Spring's MVC infrastructure, dedicated *Controller* classes acts as request handler and represent the first layer. The remaining two layers (*IVIS Mediator* and *IVIS Wrapper*) implement the Mediator-Wrapper concept, as introduced in section 2. A central *IVIS Mediator* analyses incoming requests and delegates data retrieval or modification to the appropriate *IVIS Wrapper* components. Each IVIS Wrapper class is in control of exactly one data source. Within the presented prototype, three heterogeneous data sources are connected to the system, including a relational PostgreSQL database as well as the file-based data formats XML and CSV. However, due to the generic design of the IVIS infrastructure, any additional data format can be supported, if an appropriate implementation of *IVIS Wrapper* is added to the infrastructure. If the client requests a visualization, the *IVIS Mediator* returns the queried data to the *Controller*. In accordance to the visualization pipeline, the queried data needs to be transformed from abstract data objects to visualization objects (e.g. X3DOM nodes). This transformation is executed by an *Application Template* component. To preserve universality, clients may select a certain *Application Template* through a unique identifier. In consequence, different *Application Templates* produce different visual outputs. To return the visual output or to send other messages to one or more clients, Spring's built-in *STOMP broker* is employed, offering dedicated STOMP endpoints, to which clients subscribe on application start. As each STOMP endpoint is associated with a certain functional feature of the IVIS infrastructure, clients have to subscribe to multiple endpoints. To broadcast certain messages to certain clients, the *Controller* components may send one single message to a dedicated endpoint/channel of the *STOMP broker*, who in turn delivers the message to all subscribed clients. Finally, the component *Data Source Watchers* comprises components that watch the data sources for changes. This is an important feature to detect changes within the data sources and trigger an associated client broadcast to notify them of the changed item. As response, all clients that need to update the affected data instance within their local view send an individual synchronization request to retrieve the modified object.

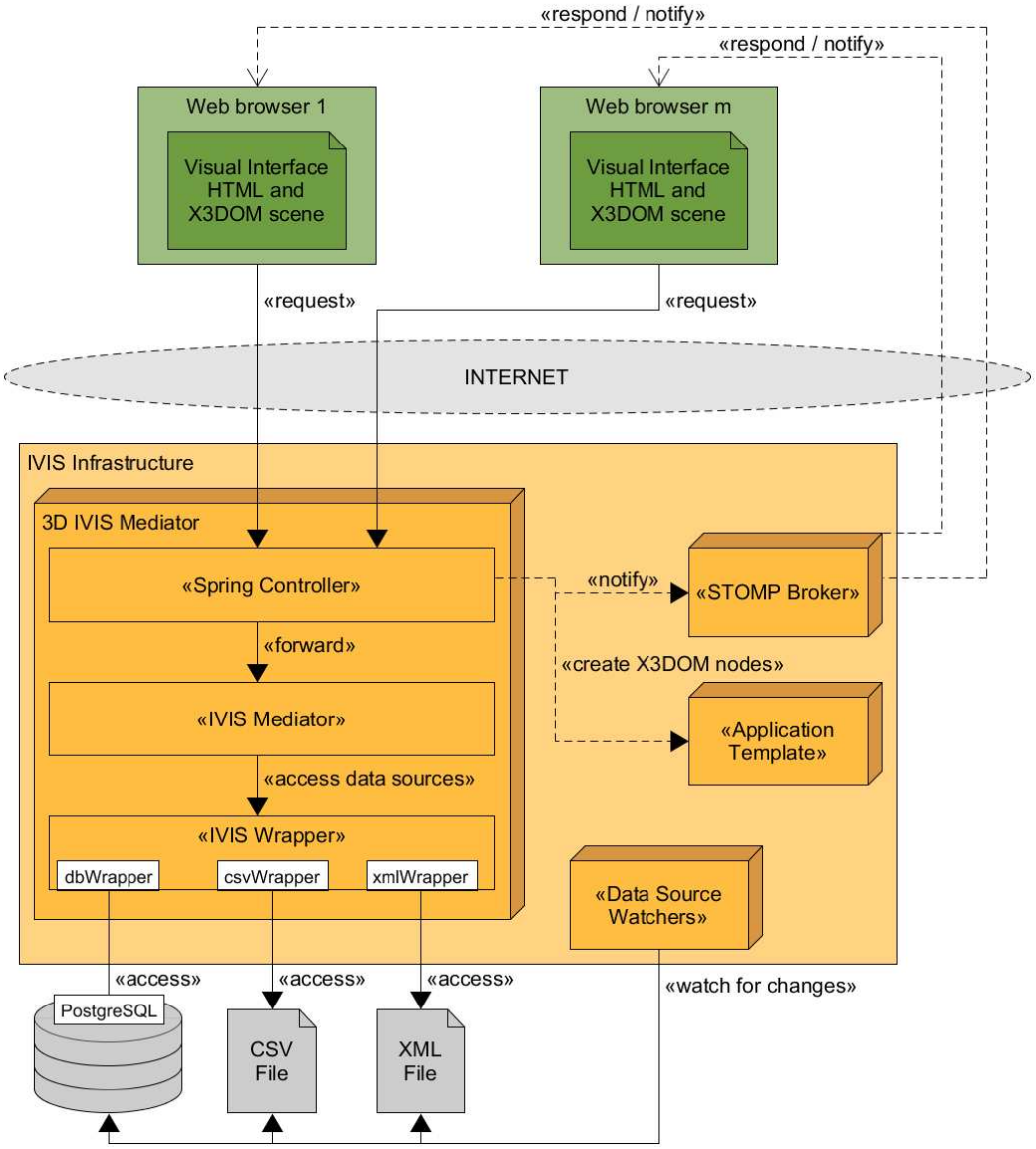


Figure 4. System Architecture of the IVIS infrastructure prototype

4.3 Semantic Integration – Implementation of the Mediator-Wrapper Pattern

As a vital infrastructure aspect, this section introduces how the semantic integration of heterogeneous data sources is implemented. Through the specification of three distinct types of mapping files, arbitrary application data from different data sources can be accessed in a homogeneous manner. For instance, concerning an online book stock information system, relevant book data may be distributed over multiple data sources. In particular, each data source may store the book properties using different syntax and/or semantics. To conquer this challenge, the mapping definition is based on a global XML schema describing the unified structure of the domain data. To retrieve certain information, clients may include a query against the global XML schema (*global query*), e.g. by including an XPath expression selecting the desired element. According to the Mediator-Wrapper pattern, the IVIS infrastructure has to solve the tasks of *creating subqueries for the affected wrappers and transform the global query into local equivalents* for the respective target data sources. To accomplish these tasks, three dedicated types of mapping files have to be specified. Each of the mapping files is presented subsequently:

1. **Global Selector to Wrapper mapping**: An XPath selector/expression against an element of the global XML schema is mapped to a list of all wrapper implementations whose data source offers associated datasets. Hence, one XPath selector is mapped to multiple wrappers (**1 global XPath Selector -> n Wrappers**).
2. **Subqueries for an aggregated Object**: Clients only include one single XPath selector within their query. However, the selected element may be an aggregated object, which is composed of several atomic properties. In consequence, this mapping file maps one XPath expression against the global schema to multiple XPath expressions against the global schema, each pointing to an atomic property of the requested element (**1 global XPath Selector -> n global XPath Selectors**). For instance, clients may query book data and thus send an XPath expression pointing to the aggregated element *book*, which consists of several properties, such as *title, author, price, stock* and *reorder* information. As a result of the mapping step, for each of the atomic properties, a new XPath expression against the global XML schema is generated and delegated to the identified wrappers from step 1.
3. **Global Selector to Local Selector (one file per wrapper)**: As final mapping step, each *global selector* against the global XML schema has to be mapped to a *local selector* using the *local syntax* and *semantics* of each wrapper's data source (**m global XPath Selector -> m local Selectors per wrapper**). E.g., if the data source is an XML file, the local selectors might be XPath expressions with different semantics. In case of a relational database, the local selector should contain information on the table and column, from where the wrapper may extract the data.

With the help of the introduced mapping mechanism, the IVIS infrastructure is kept generic to support arbitrary domain data stored in arbitrary data sources.

5 Evaluation and Exemplar Application Scenario

For demonstration purposes, the core functional features of the presented bidirectional IVIS infrastructure are presented within an exemplar bookstore application, where clients are enabled to visualize the stock value of book data as coloured extruded bars. As shown in Figure 5, clients may select certain book data by using the book properties author and price. The associated visual representations are embedded within an X3DOM scene on the right. Each book is represented by three visual elements, a text label, a thin black marker plate indicating the reorder value and a coloured extruded bar showing the stock value and reorder status (green = stock above reorder value; red = stock below reorder value and not yet reordered; blue stock below reorder value and reordered).

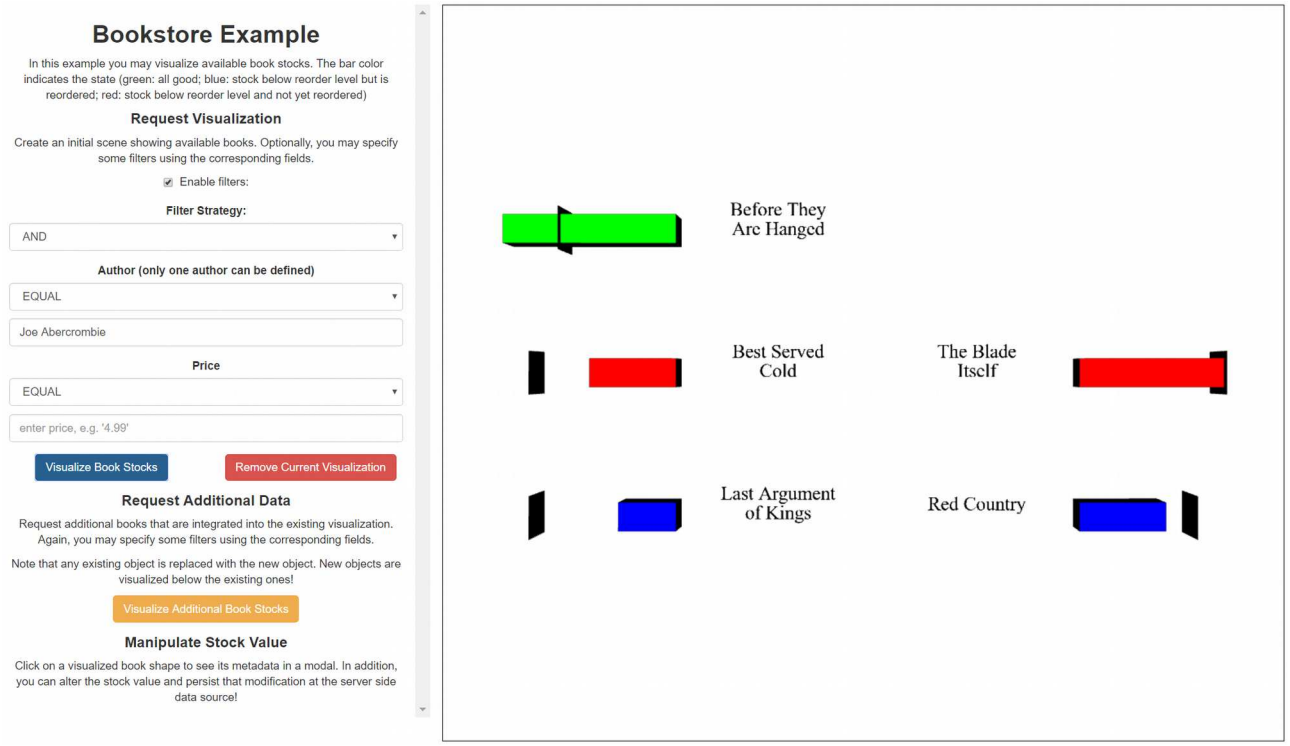


Figure 5. Bookstore Application showing visualized book stocks as coloured extruded bars

Clicking on any visualized book element, clients may enter a new stock value for the clicked book instance. The new value is transferred to and persisted within the server-side data sources. In consequence, the IVIS infrastructure broadcasts a notification to all connected clients informing them about the modified book instance. Each client, who currently visualizes the affected book instance, responds with a synchronization request to retrieve the modified book element and replace the outdated replica within the X3DOM scene. As an example, in Figure 5 the book instance with title “Last Argument of Kings” represents a stock value of 2. When it is updated to 7, the visualization is updated according to Figure 6.

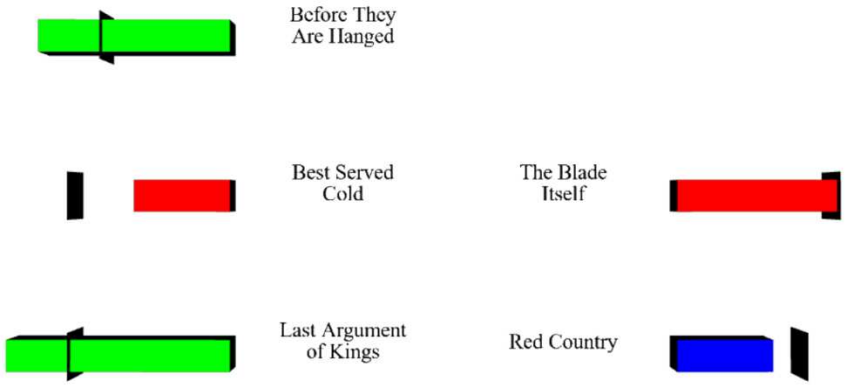


Figure 6. Synchronized visualization where the stock value of entry with title “Last Argument of Kings” has been increased from 2 to 7

6 Conclusion and Future Work

In summary, this paper presents an approach to conceptualize and implement a bidirectional Web-based 3D Information Visualization infrastructure supporting arbitrary domain data from

heterogeneous data sources, based on the Mediator-Wrapper pattern. As core technologies, the WebSocket protocol and X3DOM are utilized to establish a synchronization-enabled visual user interface allowing client modifications of data instances through the visual interface. The prototypical implementation of the server-side application logic is built using the open-source Spring framework.

With regard to future tasks, it should be noted that the presented approach represents a first step towards the generic synchronization of 3D Information Visualization and heterogeneous data sources. The implemented prototype requires more evaluation in more distinct application scenarios to improve the overall concept as well as implementation details, such as the mediation mechanism. Moreover, the support for large datasets was not focused within the initial prototype and thus has to be investigated.

In addition, the presented approach only considered synchronous use cases, while asynchronous scenarios were neglected. However, the latter reveal potential benefits concerning archiving and reproducibility. For instance, when clients perform certain interactions with the visual representation of datasets, they might want to archive a certain state including dedicated visualized instances. Hence, an export mechanism to create a permanent copy of the visualization is aspirational. A real benefit would be to store necessary information about reconnecting to the IVIS infrastructure together with the archived visualization. One the one hand, clients can then refer to the archived visualization showing the state of data instances at a certain point in time. On the other hand, if that information is attached with metadata on how to reconnect to the IVIS infrastructure (e.g. by embedding/opening the visualization within a Web browser), user will have the opportunity to re-establish that connection and receive the latest state of available data instances to update the selected elements. Of course this implies that the IVIS infrastructure including the used data instances still exists. Via such an asynchronous archiving/reconnect mechanism, users are allowed to persist a certain state of visualization for reproducibility purposes. An exemplar scenario could be to embed a visualization within a publication as picture with included metadata on how to reconnect to the IVIS infrastructure. When presenting the publication, the picture might be opened using a Web browser to automatically contact the IVIS application and retrieve the latest updates of the visualized data instances.

Another aspect worth of being investigated is how to realize additional interfaces for the respective stages of the extended IVIS Reference Model. In the proposed approach, the Web browser includes/represents the visual interface to the *View*. Concerning the remaining stages of the visualization pipeline (Raw Data, Structured Data, Visual Structures), the IVIS infrastructure could be extended to offer adequate user interfaces for each of the individual transformation stages. Thus, the respective data stages become accessible to associated specialists, as indicated by Bornschlegl et al. (2016). E.g., while *data specialists* may work directly on the (*raw* or *structured*) *data*, *Visual Analytics experts* and *Business experts* may use final *Views* to explore and analyse it. Via a separate user interface, Visual Mapping specialists can be enabled to alter the visual mapping of abstract data properties to visual attributes. In summary, specialized user interfaces for each stage of the visualization pipeline may offer adequate access for associated user stereotypes.

7 References

J. Behr, P. Eschler, Y. Jung and M. Zöllner (2009), "X3DOM: a DOM-based HTML5/X3D integration model", in: *Proceedings of the 14th International Conference on 3D Web Technology*, ACM, 2009; pp. 127-135.

M. X. Bornschlegl, K. Berwind, M. Kaufmann, F. Engel, P. Walsh, M. L. Hemmje and R. Riestra (2016), "IVIS4BigData: A Reference Model for Advanced Visual Interfaces Supporting Big Data Analysis in Virtual Research Environments", *Advanced Visual Interfaces 2016*, AVI 2016.

O. Campesato and K. Nilson (2010), *Web 2. 0 Fundamentals: with AJAX, Development Tools, and Mobile Platforms*, Jones & Bartlett Learning.

S. Card, J. Mackinlay and B. Shneiderman (1999), *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann Publishers.

C. Danowski-Buhren, M. X. Bornschlegl, B. Schmidt and M. L. Hemmje (2016), "Towards Synchronizing Data Sources and Information Visualization in Virtual Research Environments", *Advanced Visual Interfaces 2016 - Workshop "Road Mapping Infrastructures for Advanced Visual Interfaces Supporting Big Data Applications in Virtual Research Environments"*, Springer LNCS, forthcoming.

Fraunhofer-Gesellschaft (2016), "X3DOM - Instant 3D the HTML way",http://www.x3dom.org/ (Accessed 22.08.2016)

M. Kaufmann (2016), "Towards a reference model for big data management". *Research Report. Faculty of Mathematics and Computer Science*, University of Hagen.

J. Mesnil and J. Lindsay (2015), "STOMP.js", https://github.com/jmesnil/stomp-websocket (Accessed 22.08.2016)

C. North, A. Endert, C. Andrews and G. Fink (2010). "The Visualization Pipeline is Broken", FODAVA (Foundation of Data and Visual Analytics).

OECD (2007), *Participative Web and User-Created Content Web 2.0, Wikis and Social Networking*, OECD Publishing.

Pivotal Software (2016). "WebSocket Support". http://docs.spring.io/spring/docs/current/spring-framework-reference/html/websocket.html (Accessed 22.08.2016)

SockJS (2016), "SockJS-client", https://github.com/sockjs/sockjs-client, (Accessed 22.08.2016)

R. Spence (2014), *Information Visualization: An Introduction*, Springer, Berlin/Heidelberg

R. Stoyanchev (2012), "Spring MVC 3.2 Preview: Techniques for Real-time Updates", https://spring.io/blog/2012/05/08/spring-mvc-3-2-preview-techniques-for-real-time-updates/ (Accessed 22.08.2016)

V. Wang, F. Salim and P. Moskovits (2013), *The Definitive Guide to HTML5 WebSocket*, Apress.

Web3D Consortium (2016), "Why Use X3D", http://www.web3d.org/x3d/why-use-x3d (Accessed 29.06.2016)

G. Wiederhold (1992), "Mediators in the Architecture of Future Information Systems", *IEEE Computer Magazine*, Issue 04/1992, pp. 38-49.

W. Wiza (2012), "Interactive 3D Visualization of Search Results". In: W. Cellary & K. Walczak, eds. *Interactive 3D Mutlimedia Content*. London: Springer-Verlag, pp. 253-291.

Comparison of Machine Learning Algorithms in Classifying Segmented Photographs of Food for Food Logging

Patrick McAllister[1], *Huiru Zheng[1], Raymond Bond[1], Anne Moorhead[2]
[1] School of Computing & Engineering, [2] School of Communication
Ulster University, Shore Road, Newtownabbey
{mcallister-p2@email.ulster.ac.uk}, {h.zheng, rb.bond, a.moorhead}@ulster.ac.uk

Keywords: obesity, food logging, machining learning, classification

Obesity is increasing globally and is a major cause for concern (WHO, 2016). The main cause of obesity is a result of a high calorie/ fat diet and when the energy is not burned off through exercise, then much of the excess energy will be stored as fat around the body. Obesity is a serious threat to an individual's health as it can contribute to a range of major chronic conditions such as heart disease, diabetes, and some cancers (National Institutes of Health, 1998). Food logging is a popular dietary management method that has been used by individuals to monitor food intake. Food logging can include the use of text or images to document intake and research has shown that food intake monitoring can promote weight loss (Wing, 2001).

There has been much research in using computer vision algorithms to classify images of food for food logging. Computer vision methods can offer a convenient way for the user to document energy intake. The motivation for this work is to inform the development of an application that would allow users to use a polygonal tool to draw around the food item for classification. This work explores the efficacy classifying segmented items of food instead of entire food images.

This work explores machine learning (ML) techniques and feature extraction methods to classify 27 food categories with each category containing 100 segmented images. The image dataset used for this work comprises of 27 distinct food categories gathered from other research. (Jontou et al, 2009; Bossard et al, 2014). Non-food items contained in the images were removed to promote accurate feature selection (Figure 1).

Figure 1. Example of segmented food image.

Global and local feature types were extracted from the food image dataset; BoF with Speeded-Up-Robust-Features (SURF), BoF with colour features, and LBP (local binary pattern). SURF and colour features were extracted using bag of features (BoF) method to compute features for each image. A number of ML classifiers were used in this work; Sequential Minimal Optimisation (SMO, PolyKernel), Naïve Bayes (Multinomial), Neural Network (single layer, 100 nodes, 1000 epochs), and Random Forest (300 trees). Combinations of local and global features were used with ML algorithms.

Ten-fold cross validation was used to evaluate each experiment. Percentage accuracy was used to initially assess the performance of each ML algorithm. Matlab (vR2016a) was used to import and extract features from the image dataset and to export feature vectors as a CSV file. Weka (3.7.13) was used to import feature vectors and to apply ML algorithms on the feature sets. A series of classification experiments were completed using BoF with SURF and BoF with colour. The visual vocabulary used in each BoF model was changed (500 visual word increments) in each experiment to record changes in accuracy.

Table 1 and table 2 lists the results of these experiments. Further experiments were completed combining SURF and colour features. SURF and colour feature visual word sizes that achieved the highest accuracy in the previous experiments were concatenated e.g. feature length 500 achieved highest accuracy for neural network classification using colour features, and 500 for SURF, these were combined. Table 3 lists the combination percentage accuracy results.

Visual Words	Naïve Bayes (MN)	SMO	Neural Network	Random Forest
500	32.85*	42.15*	46.70*	43.33*
1000	32.07	41.07	45.48	42.26
1500	31.67	41.22	43.74	39.30
2000	31.67	41.78	43.11	38.96
2500	31.11	41.67	42.96	37.85
3000	30.89	40.89	41.56	37.22
3500	30.85	40.96	42.52	36.67
4000	30.63	40.52	39.78	35.93
4500	30.30	40.63	41.22	35.41
5000	29.85	41.70	42.04	35.41

Table 1. Percentage accuracy results of BoF with Colour features using 10-Fold cross validation.

Visual Words	Naïve Bayes (MN)	SMO	Neural Network	Random Forest
500	44.44	56.22	59.67*	46.11*
1000	45.19*	55.93	57.70	42.81
1500	45.04	56.07	57.00	42.07
2000	44.63	55.85	57.41	41.96
2500	44.33	57.15*	56.19	40.74
3000	44.48	55.89	55.81	40.07
3500	44.26	56.44	56.26	40.22
4000	43.37	56.74	56.44	39.81
4500	43.56	56.22	55.15	40.41
5000	42.96	55.51	55.74	39.00

Table 2. Percentage accuracy results of BoF with SURF features using 10-Fold cross validation.

Feature Combination	Naïve Bayes	SMO	Neural Network	Random Forest
BoF-SURF +BoF- Colour + LBP	50.48	68.29	71.77*	56.77

Table 3. Percentage accuracy results combing BoF colour and SURF features with LBP features.

The experiments focused on using an image dataset that was manually segmented to remove non-item foods or irrelevant food items from the images. Results show that using a Neural Network achieved the highest accuracy with 71.77% accuracy when combining BoF-SURF and BoF-colour features with LBP. Future work will include using other feature selection methods such as segmentation fractal texture analysis (SFTA) (Costa, 2012) gray level co-occurrence matrix (GLCM), and also exploring the use of other ML algorithms such as convolutional neural networks feature extraction and classification, and also multiclass classification methods (one vs one, one vs rest). Attribute selection methods will also be incorporated to select strongest features for classification. As well as using other ML algorithms, more food categories will be added to the food dataset.

References

Who.int. (2016). WHO | Controlling the global obesity epidemic. [online] Available at: http://www.who.int/nutrition/topics/obesity/en/ [Accessed 31 Jul. 2016].

National Institutes Of Health, 1998. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: the evidence report. Obesity Research, 6, p.51S–209S. Available at: http://www.ncbi.nlm.nih.gov/books/NBK2003/.

Wing, R.R. & Hill, J.O., 2001. Successful weight loss maintenance. Annu Rev Nutr, 21, pp.323–341. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11375440.

Taichi Joutou & Keiji Yanai, 2009. A food image recognition system with Multiple Kernel Learning. 2009 16th IEEE International Conference on Image Processing (ICIP), pp.285–288. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5413400.

Bossard, L., Guillaumin, M. & Van Gool, L., 2014. Food-101 - Mining discriminative components with random forests. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 446–461.

Costa, A.F., Humpire-Mamani, G. & Traina, A.J.M., 2012. An efficient algorithm for fractal analysis of textures. In Brazilian Symposium of Computer Graphic and Image Processing. pp. 39–46.

Chapter 5

Bioinformatics

MetaPlat: A Cloud based Platform for Analysis and Visualisation of Metagenomics Data

Nina Konstantinidiou [1], Paul Walsh [1*], Xiangwu Lu [2], Michaël Bekaert [1], Brendan Lawlor [1,2], Brian Kelly [1], Huiru Zheng [3], Fiona Browne [3], Richard Dewhurst [4], Rainer Roehe [4], Haiying Wang [3]

[1] NSilico Life Science Ltd., Cork, Ireland
[2] Department of Computer Scince, Cork Institute of Technology, Cork, Ireland
[3] Computer Science Research Institute, School of Computing and Mathematics, University of Ulster at Jordanstown, Co. Antrim, N. Ireland, BT37 0QB
[4] Future Farming Systems, Scotland's Rural College, Edinburgh, United Kingdom
[*] paul.walsh@nsilico.com

Keywords: Metagenomics, visualisation, cloud computing, sequencing, Simplicity, Bos taurus, rumen

Abstract: In recent years, the number of projects producing very large quantities of sequence data for analysis of the community structure and environmental interactions of microbes, has increased dramatically. Yet, the depth of analysis and visualisation performed is very superficial, reflecting an inefficient use of available information and financial resources. To address these deficiencies, this paper proposes the development of a fast and accurate analytic platform to fully analyse and visualise metagenomic datasets. The platform is illustrated with a use case based on metagenomic data derived from cattle (Bos taurus) intestinal digesta. By focusing on advanced hardware and software platforms, providing training and integrating diverse expertise, this project will pave the way for optimal usage of metagenomic datasets, thus reducing the number of animals required for such studies. For the livestock production industries, our approach will lead to an increase in resource use efficiency coupled with cost reductions, which will have positive effects on both financial viability and climate change consequences.

1 Introduction

Bos taurus (cattle) are ruminants, a group of mammals which also include sheep and goats. The value of domesticated ruminants comes from their ability to convert human inedible forages into high-quality, high-protein food products. This ability results from fermentation by symbiotic microorganisms in the rumen. These microorganisms confer the ability to breakdown complex polysaccharides such as cellulose and hemicellulose, that are not susceptible to mammalian enzymes. Thus, symbiosis with gut bacteria provides the host animal with nutrients, such as bacterial proteins and volatile fatty acids that are important nutrition contributors. The microbiota of the rumen, as well as other segments of the intestines, has a paramount role in cattle performance, production efficiency, immunity and health [1]. Alongside the beneficial effects of the rumen microbes, they also have a series of potential negative consequences including the biohydrogenation of polyunsaturated fatty acids (hence ruminant fat is more saturated than monogastric fat [2]) and production of methane. Ruminant methane emissions are the largest component of livestock greenhouse gas (GHG) emissions, which are estimated to make up around 11% of total anthropogenic GHG emissions [3].

Understanding the structure of the intestinal microbial community is crucial to establish links with physiological effects in the host [4]. Metagenomic sequencing, based on high-throughput analysis, offers unparalleled coverage and depth in determining the structure and dynamics of the intestinal microbial community. A key challenge, in leveraging this genomic sequencing technology, is identifying rumen microbial profiles that may be associated with phenotypic traits. Many researc8/h

groups including Mizrahi [5], Santos et al. [6], Weimer and Kohn [7], Wallace et al. [8] as well as Roehe et al. [9] have investigated the symbiotic microorganisms in the rumen because of their ecologically important traits such as feed conversion efficiency, methane production, and more recently discovery of microbes and enzymes that enable fermentation of biomass for biofuel production. A key challenge now is to identify, analyse in silico and visualise rumen microbial profiles, which are associated, and potentially predictive of these traits.

In order to build and visualise predictive models for the cattle phenotypic traits, based on rumen microbiota, both phenotype and genotype data need to be collected and cleaned. The metagenomic sequence fragments are then assembled and sorted against known annotation databases before further downstream analysis and visualisation of these data in a secure and traceable manner. Thus, to investigate microbiota in the context of genetic effects or dietary changes designed to alter methane production or feed conversion efficiency, the careful development of a sophisticated research software platform is necessary for complete analysis and visualisation of metagenomic data.

The first step, in building such a system, is the provision of a cloud-based research infrastructure that allows researchers to load and link controlled experimental data. For example, cattle identity and genotype, feed quantities and composition can be examined in the context of animal performance measurements, feed efficiency and methane output. To this end, we propose MetaPlat, a cloud-based research infrastructure for metagenomics analysis of rumen microbiota, illustrating the system that allows researchers to rapidly load, manage, analyse and visualise genome and phenome datasets along with related metadata.

In order to reduce the redundancy, the data will be characterised for assurance and will be pre-processed to remove adapters and repeating sequences using open tools. The system supports reproducibility and tracks the phenotypic information associated with each sequence, including its origin, quality, taxonomical position and any associated host genome, with the production of automated reports and visualisations of the metagenomic analyses via Simplicity's platforms. Simplicity is a powerful, yet easy-to-use bioinformatics tool which enables non-specialists to reliably run bioinformatics pipelines in a simple, safe, speedy, secure and reproducible manner.

2 Metagenomics

2.1 Introduction

In the recent years, the number of studies producing huge amounts of metagenomic data has increased dramatically [2013-2014: 182 publications, 2014-2015: 226 publications]. But the depth of analysis and visualisation is superficial, representing an inefficient use of the available information and financial resources necessary to obtain biological samples [10].

Metagenomics (also referred to as community genomics) is a useful approach for the analysis of the microorganisms that cannot be cultured in the laboratory, but which can be directly extracted from their environment. The development of metagenomics stemmed from the ineluctable evidence that as-yet-uncultured microorganisms represent the vast majority of organisms on earth. This evidence was derived from the analyses of 16S ribosomal RNA (rRNA) gene sequences amplified directly from the environment, an approach that avoids bias imposed by culturing and has led to the discovery of vast new lineages of microbial life. Although the view of the microbial world is still limited by analysis of 16S rRNA genes alone, such studies yielded a substantial advance in the

phylogenetic description and profiling of community membership. Metagenomics based on high-throughput sequencing offers unparalleled coverage and depth in determining microbial dynamics as long as the analytic resources and visualisation tools are available. These studies have been applied to a number of organisms including humans, but in this work we focus on Bos taurus.

2.2 In silico Analysis of Metagenomic Data

Careful development of a software platform with comprehensive visualisation tools for more complete in silico analyses of metagenomic data is essential for further advances in understanding of relationships between the gut microbiota, dietary changes, methane production and feed conversion efficiency. Currently, some generic sequence databases, such as DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) and National Center for Biotechnology Information (NCBI), are the only sources of reference sequences for the classification of newly identified species, based on their nucleic acid sequences. However, there has been no systematic attempt to classify uncategorised groups or reduce the level of redundancy produced, despite the general effort/wish to curate quality sequence data, which is a central objective in MetaPlat.

A related problem is the lack of new algorithms, approaches and visualisation tools (for example, comprehensive phylogenetic trees) to deal with the huge number of sequences generated via sequencing of bacterial metagenomes. Usually, assignment of a sequence to an organism is performed by sequence similarity, which is not optimal if the sequences diverge greatly or not enough. For example, if two or more sequences are identical, the sequence/species assignment is often given by the first sequence name encountered by the algorithm, which can be ambiguous and even missing, producing very biased and non-reproducible results and interpretations. In addition to lack of reproducibility and inadequate visualisation of results, significantly low speed and scalability of the analysis are now concerns for the field, not to mention the lack of skilled bioinformaticians. Thus, robust computational approaches are required to distil big data derived from metagenomics analysis down to clear visualisation of the gut microbiota community structure, while supporting and enhancing reproducibility.

2.3 Related Work

Research by Brulc et al. [11] and Hess et al. [12] were seminal in the development of rumen metagenomics, but their studies involved a few animals. Sequencing costs are reducing rapidly and the ability to sequence either multiple 16S libraries or whole rumen metagenomes has advanced to the point where data analysis capability and acquisition of phenotypes are now major constraints, rather than the cost of sequencing.

Almost all tools currently available assign a sequence to an organism by sequence similarity after sequence alignment (Table 1). This is a computationally intensive and time consuming task, and these tools are not optimised for parallel or cloud execution, nor for visualisation. Some tools also rely on scriptic languages and algorithms which are not optimised. The lack of new algorithms, optimised approaches and visualisation tools, to deal with the huge number of generated sequences, increases bias and contributes to the production of non-reproducible results [10]. Representation of results for dissemination is also an important challenge. Today, only a single tool used by almost all studies, Krona [13], provides a valuable interactive generic hierarchical data browser, but does not convey all degrees of information and visualisation.

Software	Reference	Access	Comments
RDP suit	[14]	http://rdp.cme.msu.edu/	Single sequence request, BLAST alignment based
PhyloSift	[15]	https://phylosift.wordpress.com/	NGS data, Phylgeny-driven classification
MGTAXA		http://mgtaxa.jcvi.org/	NGS data, Phylgeny-driven classification
PhymmBL	[16]	http://www.cbcb.umd.edu/software/phymmbl/	Discontinued, BLAST alignment based
MG-RAST	[17]	http://metagenomics.anl.gov/	Online only, NGS data, BLAST alignment based
Metaxa2	[18]	http://microbiology.se/software/metaxa2	Perl scripts, NGS data, BLAST alignment based
MetaPhlAn	[19]	http://huttenhower.sph.harvard.edu/metaphlan/	NGS data, BLAST alignment based
Parallel-META	[20]	http://www.computationalbioenergy.org/parallel-meta.html	NGS data, BLAST alignment based
MetaPhyl	[21]	http://www.sc.ucr.edu/~tanaseio/metaphyl.html	Training data set required, k-mer based
CLcommunity		http://www.chunlab.com/software_clcommunity_about/	Licenced, NGS data, BLAST alignment based
QIIME	[22]	http://qiime.org/	Performs microbiome analyses
Cytoscape	[23]	http://www.cytoscape.org/	Produces molecular interaction networks
STRING	[24]	http://string-db.org/	Produces functional protein association networks

Table 1. Online tools for the analysis of metagenomic data

3 Materials and Methods

To address the challenges outlined in the previous section, we will present a methodology that will interlink important research elements in the metagenomics domain: Biological samples (3.1), Big data management, characterisation and curation (3.2), Reference database distribution (3.3), In silico search algorithms (3.4), Phylogeny-aware classification (3.5), Machine learning models (3.6), New visualisations (3.7), Integrated software pipeline development using high-performance computing (HPC) platforms (3.8) and Reproducibility (3.9). The platform will be developed based on a Simplicity pipeline [25].

3.1 Biological Samples

In the first stage, biological samples will be collected and sequenced in parallel to reference databases preparation and algorithm development. Rumen fluid samples will be collected using a naso-gastric sample tube system at Scotland's Rural College (SRUC). DNA isolation will be carried

out with a bead-beating procedure according to Yu and Morrison [26] and 16S amplicons will be prepared using Caporaso primers [27]. Sequencing of both 16S amplicons (in parallel) and total genomic DNA will be accomplished using Illumina sequencers (MiSeq and HiSeq).

3.2 Big Data Management, Characterisation and Curation

Characterisation of sequenced data will facilitate better understanding of the bacterial communities collected from B. taurus rumen. Accurate classification of new sequences entails the formal and reliable analysis of the complete set of reference sequences. In order to reduce the redundancy, a system will be developed to keep track of each sequence origin, taxonomical position, associated bacterial genome (gene set), quality and redundancy status. This will involve the development of a system allowing automatised update of the databases and the public accessibility of the resulting up-to-date dataset. Another system will be designed to provide homogenised and high quality data, by analysing and visualising the results in a fully integrated manner (Figure 1).

3.3 Reference Database Distribution

To create a reference database, data from the publicly available reference databases such as GreenGenes [http://greengenes.lbl.gov/cgi-bin/nph-index.cgi] will be used in the metagenomics pipeline (Figure 1). This reference database will be updated automatically via a relevant system allowing automated update of the data. This system will be designed to provide homogenised and high quality data, by visualising and analysing the sequenced data in a fully integrated manner.

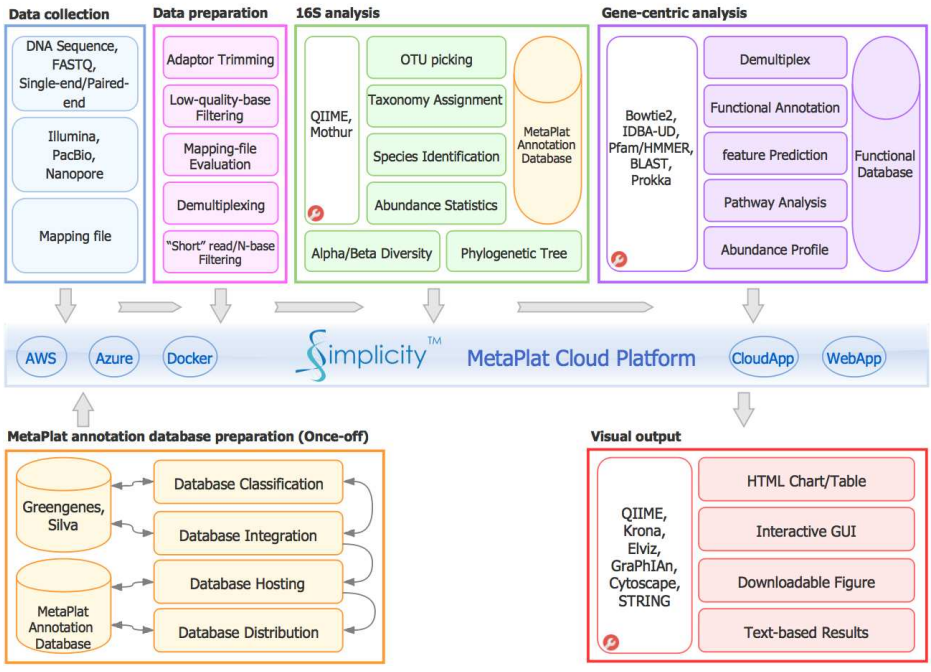


Figure 1. Illustration of a high level Big data reference model based on cloud technology. Big data and Simplicity's services and analysis will be stored in a cloud. This will reduce the required space for the analysis of large datasets by the users. Diverse visualisations will be integrated into the metagenomics pipeline of Simplicity to enhance classification and interpretation of biological data.

3.4 In silico Search Algorithms

Metagenomic samples contain reads from a huge number of organisms. For example, in a single gram of soil, there can be up to 18,000 different types of microbes, each with different genomic background. Usually, classical binning approaches are used for grouping reads or contigs and assigning them to operational taxonomic units (OTUs). Based on Mande et al. [28] review, we will develop a new algorithmic approach, facilitating accuracy, speed, reproducibility and visualisation of the analysis that will subsequently be implemented in a High-Performance Computing (HPC) infrastructure. We will build on the known and trusted classification algorithms to implement a comprehensive pipeline of algorithms for the integrated analysis and visualisation of microbial populations in Simplicity's metagenomics pipeline.

3.5 Phylogeny-aware Classification

Phylogeny awareness is an important element for the annotation and classification of the sequences. The main objective of this task is to develop a new phylogeny-aware pipeline, within Simplicity's metagenomics platform, in order to classify sequencing data. This algorithm will facilitate annotation of each sequence based on phylogenetic information available online.

3.6 Machine Learning Models

Machine learning models are useful for the prediction of previously unclassified sequences. Currently, greedy and hierachical clustering are the most frequently used machine learning algorithms for metagenomic data analysis [29]. In this study, a new machine learning based classification model will be proposed and implemented into the Simplicity's metagenomics pipeline. To train, test and evaluate this model, the curated data with phylogeny-aware classification will be utilised in the pipeline.

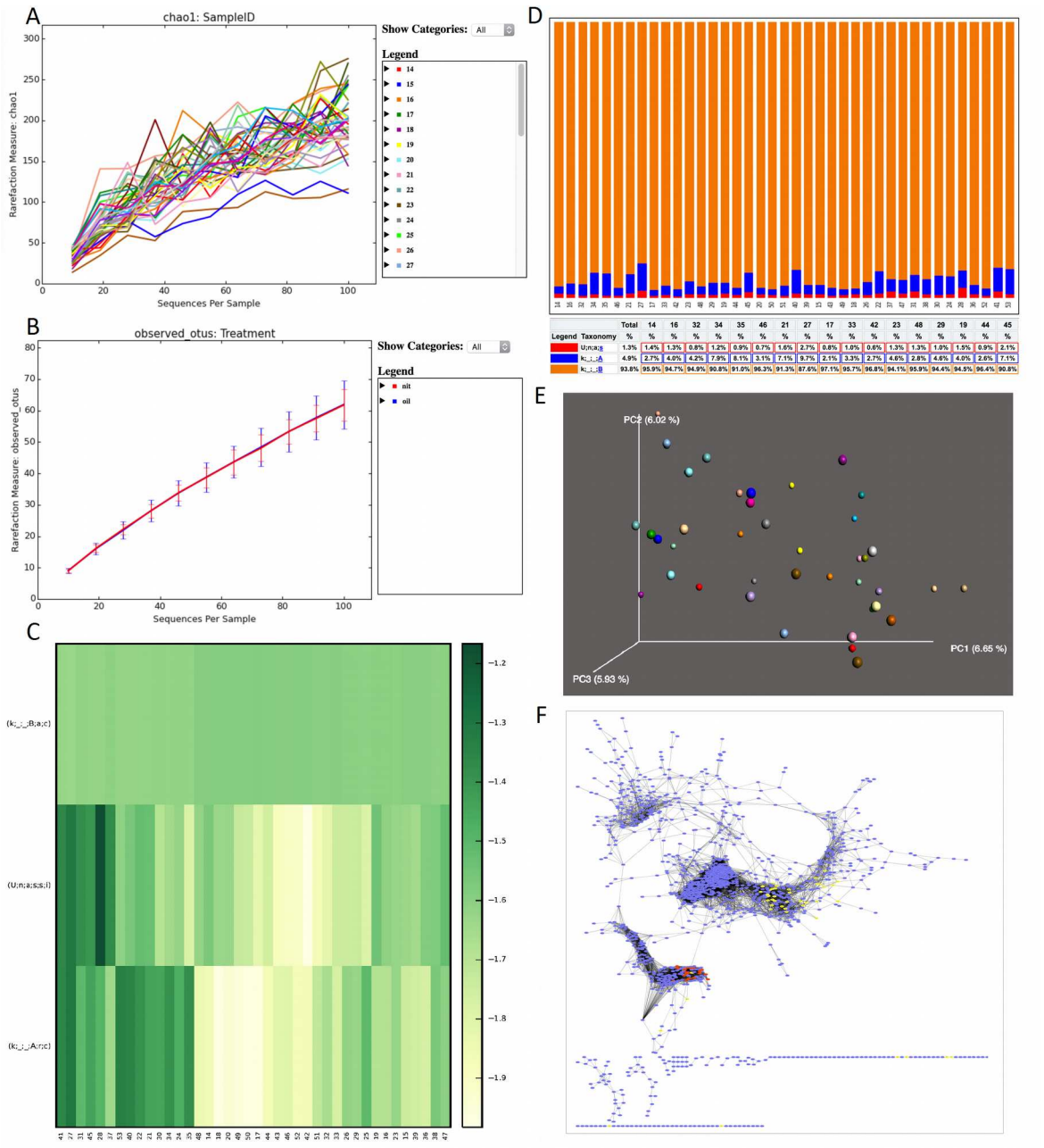


Figure 2. Preliminary QIIME analyses of a dataset from Teagasc, Ireland (A-E) and Cytoscape analysis of a dataset from Scotland's Rural College (F). (A) Alpha (α) diversity of the species richness within the community is calculated using Chao1 metric and represented as a line chart with QIIME (B) Alpha (α) diversity is measured using Shannon index. The Shannon index denotes the entropy of the observed OTU abundances and accounts for both richness and evenness. (C) Heatmap reflects the matrix of OTU abundance in each sample with meaningful taxonomic identifiers for each OTU. (D) Bar chart illustrates the taxa assignments for each sample within communities. (E) Principal Coordinates Analysis (PCoA) plot demonstrates the hierarchical clustering by transforming the distance matrix into three dimensional (3D) coordinates using beta (β) diversity metrics. The similarity of samples is shown by smaller distance between the dots. (F) Co-abundance network analysis using Cytoscape. With the threshold of 0.95, the yellow nodes represent microbial genes associated with feed conversion efficiency. Red nodes illustrate methane emission-related microbial genes which were identified in Scotland's Rural College (SRUC) data [9].

3.7 New Visualisations

Visualisation of the metagenomic data is crucial for understanding the microbial diversity in the gut. To assign biological meaning to the in silico metagenomic analysis, Krona [13] and additional visualisation tools will be integrated into the Simplicity (Figure 1). For example, QIIME (Quantitative Insights Into Microbial Ecology) [22] and Elviz [30] are designed to analysed 16S rRNA gene sequences of microbial communities and for interactive visualisation of metagenome assemblies respectively. These tools will be incorporated into the metagenomics pipeline of Simplicity to visualise microbial data through PCoA plots, bar charts (QIIME) (Figure 2) and bubble plots (Elviz). To reveal the evolutionary lineages of the microbes in the gut, phylogenetic analysis of metagenomic data needs to be conducted using additional visualisation aspects. Due to the latter, GraPhlAn [31], a compact graphical representation tool for metadata, will be integrated into the Simplicity's pipelines.

Visualisation of protein-protein interaction (PPI) networks can reveal biological functions of the proteins and the molecular processes they are involved. To carry out network analysis, Cytoscape [23] and STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins) [24] can additionally be employed for a graphical output of PPI networks and visualisation of functional protein associations respectively. For paradigm, using Cytoscape PPI network, the preliminary analysis of Scotland's Rural College (SRUC) data highlighted two distinct sets of interesting genes (Figure 2). Figure 2F shows that the microbial genes linked with feed conversion efficiency (yellow nodes) and those related with methane emission (red nodes) were clustered into two different clusters, indicating functional differences. Even though diverse visualisation tools will be added to the metagenomics pipeline, Simplicity will maintain its easy-to-use features. Importantly, the reproducibility of the analyses and visualisations will be secured via Docker technology (see below section 3.9 Reproducibility).

3.8 Integrated Software Pipeline Development using High-Performance Computing Platforms

For the development of Simplicity's metagenomics pipeline, a formal software development approach with a standard software engineering model will be applied. Software development will be initiated with careful analysis of the requirements, by engaging with multidisciplinary teams of the researchers including microbiologists, molecular biologists, clinicians, bioinformaticians and developers. The core effort will be to develop thread safe extensible bioinformatics algorithms for high-performance computing (HPC) architectures using ISO/IEC 12207, an international standard for software life-cycle processes, which defines best practices for developing and maintaining software. A process will be put in place to identify the most scalable parallel computing architectures and the appropriate functional partitioning strategies to achieve the most optimal solution by maximising data locality and minimising inter-process communication.

3.9 Reproducibility

Reproducibility is an important aspect of the research but difficult to achieve while using in silico techniques. We propose to address this using Docker technology [32], which will facilitate reproducibility by encapsulating a complete environment with scripts, libraries, databases, system tools and tool dependencies. This approach will provide reproducibility and transparency in experimental methodology, observation and collection of data through the development of web-

based tools by facilitating collaboration, storage and integration of new data and visualisations. Data will be shared and archived using digital information objects interlinked with internal and external resources, in a structured and machine readable serialization mechanism, as measured against ISO 14721:2012 (OAIS), by adding the required provenance and descriptive. For instance, domain information (metadata) for ensuring results may be shared and reproduced over long term data life cycles. In order to support digital preservation by paving the way towards reproducibility, previous results on reproducibility of bioinformatics analyses by means of archiving research data in archival packages, are compliant with ISO 14721:2012 (Open Archival Information System – OAIS).

4 Discussion, Summary and Future Work

In this paper we have highlighted key challenges to be addressed in the metagenomics domain and highlighted our proposed solution, MetaPlat. MetaPlat will integrate complementary innovative activities, including new visualisation options, to address the limitations in previous work, and thus underline the overall metagenomics pipeline development plan within Simplicity. Currently, integration and evaluation of metagenomics analysis and visualisation tools has been initiated in collaboration with microbiologists and bioinformaticians. Thus, the following innovations, outlining metagenomics data analysis pipeline, are noteworthy:

Development of non-redundant sequence databases, that will encapsulate all currently available good quality metadata (species and sub-species/strain, number of sequence), as well as all previously unclassified sequences, using a machine learning approach. This will include a procedure to update the data with newly released information and a public dissemination of the resulting database. This database will be incorporated within Simplicity's metagenomics pipeline, a new platform focusing on reproducible metagenomics analysis in silico.

Development of accurate and reproducible classification algorithms, optimised for parallel and cloud execution with open source code. These algorithms will be organised into services and pipelines, collectively leading to the development of Simplicity's metagenomics pipeline, a revolutionary bioinformatics tool for in silico analysis of complex metagenomic data.

Real-time or time-efficient comparisons and analyses of large datasets with phylogeny-aware classification, but also quantitative and functional analyses (when possible) will also be carried out using Simplicity's metagenomics comparison pipeline that is able to compare hundreds of bacteria.

Production of statistical and visual representations in a standard report, after data analysis with Simplicity's metagenomics pipeline, can maximise microbiome information and convey biological meaning to current visualisations.

Simplicity's metagenomics pipeline will be able to perform standard steps of high quality analyses and comparisons of public databases as well as user-own dataset with a few clicks automatically, in an integrated environment (e.g. Galaxy [33] web-based platform).

It is envisaged, that the Simplicity metagenomics pipeline output, addressing key objectives of the MetaPlat project, will revolutionize gut microbiome analysis. This is important for advancing our understanding of gut microbiota to benefit humans, animals and the environment. These benefits include improved animal health and productivity, reduced methane emissions and improved targeting of dietary supplements. Future work will focus on the development of machine learning

and visualisation models for metagenomics analysis in order to integrate them into the next versions of Simplicity's metagenomics pipeline.

5 Acknowledgements

This research is supported by Horizon 2020 MSCA-RISE-2015 Marie Skłodowska-Curie Research and Innovation Staff Exchange Project (ID: 690998), www.metaplat.eu.

6 References

1. Beever, D.E.: The Impact of Controlled Nutrition During the Dry Period on Dairy Cow Health, Fertility and Performance. Anim Reprod Sci. 96, 212-226 (2006).

2. Dewhurst, R.J.: Targets for Milk Fat Research: Nutrient, Nuisance or Nutraceutical? Journal of Agricultural Science 143, 359-367 (2005).

3. Gerber, P.J., Hristov, A.N., Henderson, B., Makkar, H., Oh, J., Lee, C., Meinen, R., Montes, F., Ott, T., Firkins, J., Rotz, A., Dell, C., Adesogan, A.T., Yang, W.Z., Tricarico, J.M., Kebreab, E., Waghorn, G., Dijkstra, J., Oosting, S.: Technical Options for the Mitigation of Direct Methane and Nitrous Oxide Emissions from Livestock: a Review. Animal 7, 220-234 (2013).

4. Oskoueian, E., Abdullah, N., Oskoueian, A.: Effects of Flavonoids on Rumen Fermentation Activity, Methane Production, and Microbial Population. Biomed Res Int. 2013, 349129 (2013).

5. Mizrahi, I.: The Role of the Rumen Microbiota in Determining the Feed Efficiency of Dairy Cows. Springer 14, 203-210 (2011).

6. Santos, V.C. , Ezequiel, J.M.B. , Homem Junior, A.C. , Pinheiro, R.S.B.: Quantification of Ruminal Microbiota and Production of Methane and Aarbonic Dioxide from Diets with Inclusion of Glycerin. Arq Bras Med Vet Zootec. 67, 1678-4162 (2015).

7. Weimer, P.J., Kohn, R.A.: Impacts of Ruminal Microorganisms on the Production of Huels: How Can We Intercede from the Outside? Appl Microbiol Biotechnol. 100, 3389-98 (2016).

8. Wallace, R.J., Rooke, J.A., McKain, N., Duthie, C-A., Hyslop, J.J., Ross, D.W., Waterhouse, A., Watson, M., Roehe, R.: The Rumen Microbial Metagenome Associated with High Methane Production in Cattle. BMC Genomics 16, 839 (2015).

9. Roehe, R., Dewhurst, R., Duthie, C-A., Rooke, J.A., McKain, N., Ross, D.W., Hyslop, J.J., Waterhouse, A., Watson, M., Wallace, R.J.: Bovine Host Genetic Variation Influences Rumen Microbial Methane Production with Best Selection Criterion for Low Methane Emitting and Efficiently Feed Converting Hosts Based on Metagenomic Gene Abundance. PLoS Genetics 12, e1005846 (2016).

10. Ioannidis, J.P.A.: Why Most Published Research Findings Are False. PLoS Med. 2, e124 (2005).

11. Brulc, J.M., Antonopoulos, D.A., Miller, M.E., Wilson, M.K., Yannarell, A.C., Dinsdale, E.A., Edwards, R.E., Frank, E.D., Emerson, J.B., Wacklin, P., Coutinho, P.M., Henrissat, B., Nelson, K.E., White, B.A.: Gene-centric Metagenomics of the Fiber-adherent Bovine Rumen Microbiome Reveals Forage Specific Glycoside Hydrolases. Proc Natl Acad Sci USA 106, 1948-1953 (2009).

12. Hess, M., Sczyrba, A., Egan, R., Kim, T.W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., Mackie, R.I., Pennacchio, L.A., Tringe, S.G., Visel, A., Woyke, T., Wang, Z., Rubin, E.M.: Metagenomic Discovery of Biomassdegrading Genes and Genomes from Cow Rumen. Science 331, 463-467 (201).

13. Ondov, B.D., Bergman, N.H., Phillippy, A.M.: Interactive Metagenomic Visualization in a Web Browser. BMC Bioinformatics 12, 385 (2011).

14. Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R.: Naive Bayesian Classifier for Rapid Assignment of RRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 73, 5261–5267 (2007).

15. Darling, A.E., Jospin, G., Lowe, E., Matsen IV, F.A., Bik, H.M., Eisen, J.A.: PhyloSift: Phylogenetic Analysis of Genomes and Metagenomes. PeerJ. 2, e243 (2014).

16. Brady, A., Salzberg, S.: PhymmBL Expanded: Confidence Scores, Custom Databases, Parallelization and More. Nat Methods 8, 367 (2011).

17. Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., Edwards, R.A.: The Metagenomics RAST Server – a Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes. BMC Bioinformatics 9, 386 (2008).

18. Bengtsson-Palme, J., Hartmann, M., Eriksson, K.M., Pal, C., Thorell, K., Larsson, D.G.J., Nilsson, R.H.: METAXA2: Improved Identification and Taxonomic Classification of Small and Large Subunit rRNA in Metagenomic Data. Mol Ecol Resour. 15, 1403-1414 (2015).

19. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., Huttenhower, C.: Metagenomic Microbial Community Profiling Using Unique Clade-Specific Marker Genes. Nat Methods 9, 811–814 (2012).

20. Su, X., Pan, W., Song, B., Xu, J., Ning, K.: Parallel-META 2.0: enhanced Metagenomic Data Analysis with Functional Annotation, High Performance Computing and Advanced Visualization. PLoS One 9, e89323 (2014).

21. Tanaseichuk, O., Borneman, J., Jiang, T.: Phylogeny-based Classification of Microbial Communities. Bioinformatics 30, 449-456 (2014).

22. Kuczynski, J., Stombaugh, J., Walters, W.A., González, A., Caporaso, J.G., Knight, R.: Using QIIME to analyze 16S rRNA gene sequences from Microbial Communities. Curr Protoc in Bioinformatics, Chapter 10, Unit 10.7 (2011).

23. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 13, 2498-2504 (2003).

24. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., von Mering, C.: STRING 8--a Global View on Proteins and Their Functional Interactions in 630 Organisms. Nucleic Acids Res. 37, D412-416 (2009).

25. NSilico Life Science Ltd.: SimplicityTM Version 1.5, Cork, Ireland (2015).

26. Yu, Z., Morrison, M.: Improved Extraction of PCR-quality Community DNA from Digesta and Fecal Samples. Biotechniques 36, 808-812 (2004).

27. Caporaso, J., Lauber, C., Walters, W., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J., Smith, G., Knight, R.: Ultrahigh-throughput Microbial Community Analysis on the Illumina HiSeq and MiSeq Platforms. ISME J. 6, 1621-1624 (2012).

28. Mande, S.S., Mohammed, M.H., Ghosh, T.S.: Classification of Metagenomic Sequences: Methods and Challenges. Brief Bioinform. 13, 669-681 (2012).

29. Soueidan, H., Nikolski, M.: Machine Learning for Metagenomics: Methods and Tools. Review Metagenomics 1, 1–19 (2016).

30. Cantor, M., Nordberg, H., Smirnova, T., Hess, M., Tringe, S., Dubchak, I.: Elviz - Exploration of Metagenome Assemblies with an Interactive Visualization Tool. BMC Bioinformatics 16, 130 (2015).

31. Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., Segata, N.: Compact Graphical Representation of Phylogenetic Data and Metadata with GraPhlAn. PeerJ. 3, e1029 (2015).

32. Boettiger, C.: An Introduction to Docker for Reproducible Research. ACM SIGOPS Operating Systems Review 49, 71-79 (2015).

33. Goecks, J., Nekrutenko, A., Taylor, J., Galaxy Team: Galaxy: a Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences. Genome Biol. 11, R86 (2010).

Towards a host-pathogen integrated molecular diagnostic for bacterial infection in newborn babies

Forster T [1,2], Dantoft W [1,2], Kropp K [6], Dickinson P [1], Smith CE [5], Mullins D [3,4], Kelly B [2], Sleator R [4], Lawlor B [2], Konstantinidiou N [2], Walsh P [2], Ghazal P [1]

[1] Division of Infection and Pathway Medicine, Edinburgh Medical School, Edinburgh, United Kingdom

[2] NSilico Life Science Ltd., Cork

[3] School of Microbiology, University College Cork, Cork, Ireland

[4] Department of Biological Sciences, Cork Institute of Technology, Bishopstown, Cork, Ireland

[5] Neonatal Unit, Simpson Centre for Reproductive Health, Royal Infirmary of Edinburgh

[6] Institute of Virology, Hannover Medical School, Hannover, Germany

p.ghazal@ed.ac.uk

Keywords

Bacterial infection, sepsis, classification, host immune response, microarray, sequencing, software

Abstract

Sepsis is a life-threatening immune response to bacterial infection that is difficult to accurately diagnose. Little is known about the human neonatal immune system and in clinical practice, there is a low threshold to treat with broad-spectrum antibiotics. This poses complications toward stewardship of antimicrobial resistance and can have iatrogenic effects.

We hypothesized that the human host immune response (at the level of transcribed genes) integrated with genomic data of the infecting microbes can be used to develop a quick and reliable diagnostic assay that identifies if a bacterial infection is present in a blood sample.

To develop this approach beyond the identification of a suitable set of genes for diagnosing bacterial infection, translating this into an assay and clinical use requires interdisciplinary expertise in clinical medicine, host and pathogen biology, data science and computing software and infrastructure. To address this in absence of standardised ways of accomplishing such goals, the ClouDx-i (EU FP7 IAPP) project has enabled support for a multidisciplinary approach. ClouDx-i employs a suite of biomedical and computational approaches to address the hypothesis, this includes microarray transcription arrays, next generation sequencing, and multiple analytical and computational solutions.

We have identified and independently validated a promising host-based classifier gene set for clinical diagnosis of neonates presenting with symptoms of infection and early signs of sepsis. We have also identified laboratory, analysis and computational processes that aid in the genomic identification of the microbial agent. In the course of our studies we identified process bottlenecks and cross-disciplinary obstacles relevant to driving the research life-cycle from hypothesis to clinical implementation. Our ultimate research objectives are to (a) undertake a neonatal multi-centre study to evaluate the clinical use of this classifier, and (b) to promote academic-industry collaboration for establishing a methodological and software pipeline that can be used for generating diagnostic tests for other human diseases.

1 Introduction

Sepsis is an uncontrollable escalation of the immune system in response to infection or injury (Cohen, 2002), resulting in complications, morbidity and mortality. This is a particular concern for neonates (newborn babies), as of 2010, ~3.1 million neonatal deaths could be attributed to infectious causes (Liu et al, 2012), and within the population of neonates up to 65% of those with very low birthweights develop presumed sepsis (Stoll et al, 2004). For clinicians, rapid diagnosis and appropriate treatment are paramount in neonates, and current practice entails broad-spectrum antibiotics given on suspicion of a diagnosis of sepsis. However, diagnosis of bacterial sepsis is

problematic in that clinical presentation of sepsis varies from harmless or non-specific to dramatic symptoms, and diagnostics assays for the source of infections are both slow and not highly accurate. Since the risk of missing a diagnosis of bacterial sepsis can quickly lead to severe consequences, many infants without any infection or any bacterial infection are exposed to broad-spectrum antibiotic treatment that in these cases is not indicated and can have detrimental side-effects. It is therefore highly desirable to identify means for a fast and reliable diagnosis of the presence of bacterial infectious agents at time of presentation in order for clinicians to decide on treatment (no treatment, specific antibiotic, broad-spectrum antibiotic).

Lacking in this context is both understanding of infections in neonates, and absence of animal models to study biological mechanisms of neonatal infection. The importance of the host response to pathogens has been noted (Manger & Relman, 2000), as has the fact that this host response can now be routinely measured at the gene level by genomic technology platforms (Hyatt et al, 2006). Multiple efforts have been undertaken to identify biomarkers related to neonatal sepsis (Andrade et al, 2008; Kingsmore et al, 2008; Wong, 2013; Wong et al, 2009a; Wong et al, 2009b). However, identification of a list of bacterial sepsis biomarkers using high-throughput, highly parallel laboratory techniques like microarrays differs from providing a diagnostic tool for clinical practice and from an understanding of the biological mechanisms of the host response as well as host-pathogen interactions during infection. We therefore set out our hypothesis that a molecular classifier can be used to identify bacterial infection in whole blood samples from neonates at clinical presentation. We further proposed that this needs be underpinned by sequencing pathogens isolated from individual patients in order to obtain information on bacterial strain, antimicrobial resistance genes or virulence genes. We propose that this process of developing such a tool is complex and requires the collaboration of multiple disciplines that cover clinical, biological, and data science expertise. We adapted this interdisciplinary process into ClouDx-I (Cloud Based Software Solution for Next Generation Diagnostics in Infectious Diseases), a European Union Framework 7 application to the Marie Curie Industry-Academia Partnership and Pathways (CFP7-PEOPLE-2012-IAPP), linking industry computer science and bioinformatics experience to academic base research and clinical studies. Towards the objective of a molecular classifier for bacterial sepsis in neonates, we utilise a clinical patient study, identify a host-centric classifier gene set, analyse pathogen genomes, devise an extended classifier validation strategy and plan further hypothesis-testing studies closer to clinical application of this classifier as a diagnostic tool. The participants in this are the Cork Institute of Technology (CIT, Ireland), NSilico Life Science Ltd (NS, Ireland) and the University of Edinburgh Division of Infection and Pathway Medicine (DIPM, UK). Figure 1 below details our work to date and subsequent stages.

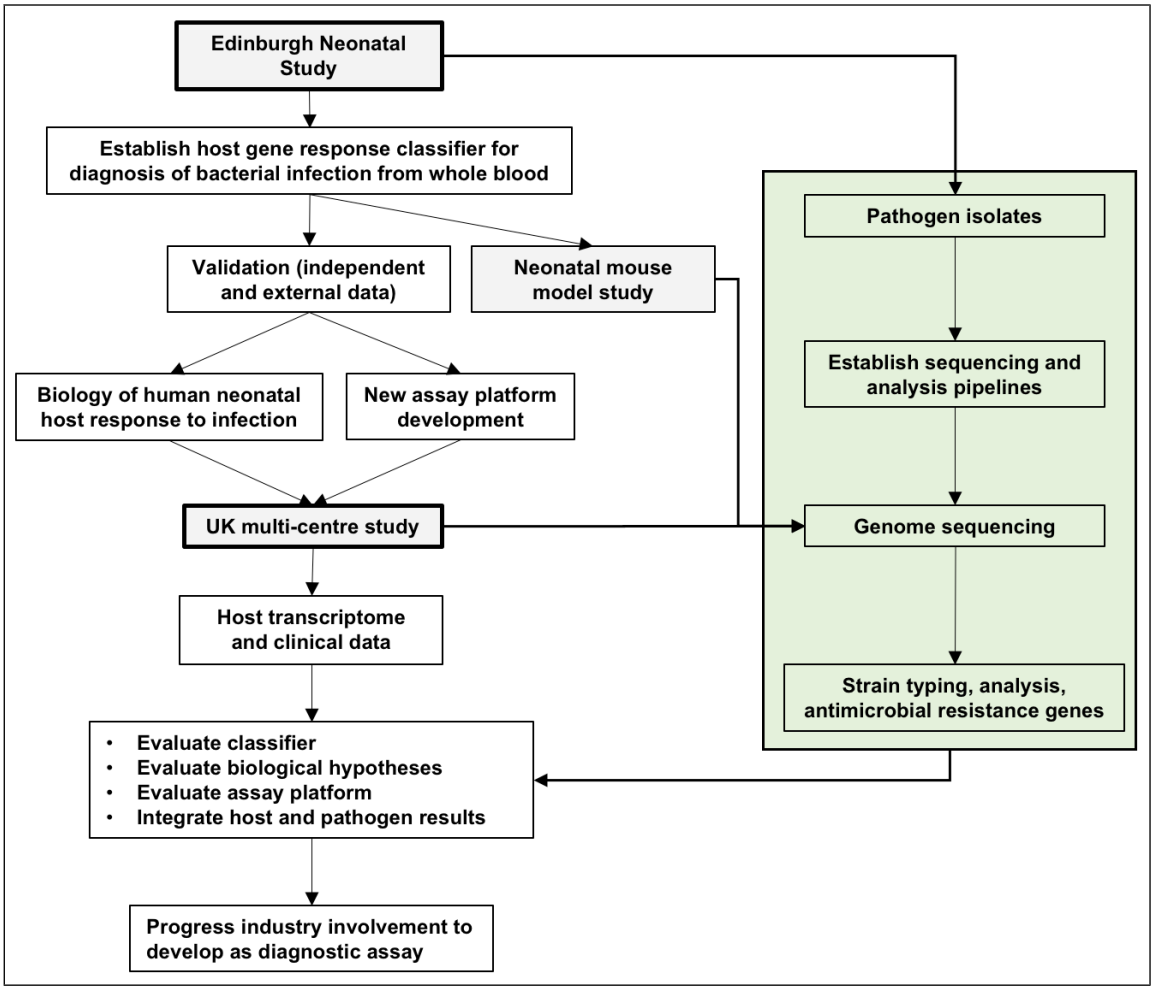
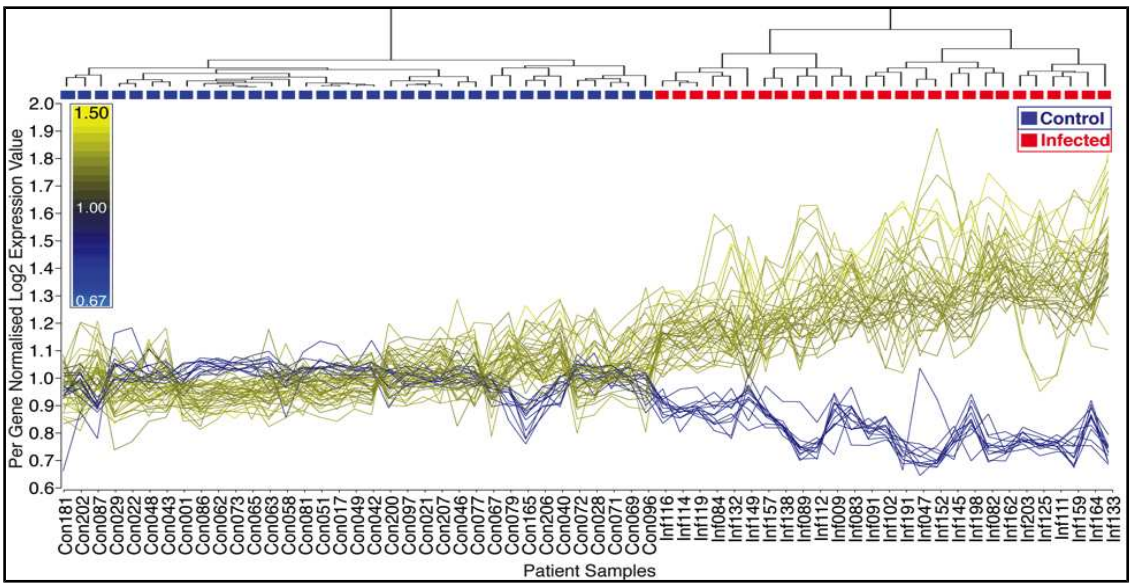


Figure 1. Towards a molecular diagnostic for neonatal sepsis

1.1 Classifier for predicting presence of bacterial infection from host gene markers

We initially carried out a clinical study collecting blood samples of 35 bacterially infected neonates (displaying signs of sepsis) and 27 control neonates at the Royal Infirmary of Edinburgh Neonatal Unit (Smith et al, 2014). Using microarray transcription array technology, we profiled the transcriptome of each individual sample to determine the activity level ('expression') of each known human gene. After initial statistical analysis of differential gene expression between control neonates and bacterial infection neonates we identified classification features (also known as predictors or in this case, genes) by way of a pathway biology led approach, removing statistically significant genes that were associated with blood development and retaining statistically significant genes that were associated with either human innate immunity, adaptive immunity or metabolism. Figure 1 shows the expression profile of 52 genes across all samples, demonstrating that with reference to control samples both up and down regulation are features of the host transcriptional response.



Neonatal samples in x-axis, standardised (mean=0, sd=1) gene expression on y-axis. One line per gene probe, coloured to indicate up or down regulation between control and infected conditions.
Colour bar at top represents sample type for the imposed hierarchical clustering
(dendrogram at top). (Smith et al, 2014), Figure 2:

Figure 2 - 52-Gene dual-network classifier of neonatal bacterial infection comprised of innate, metabolic and adaptive immune pathways.

We trained a Receiver-Operating-Characteristic (ROC) classification algorithm on the selected 52-gene set in this study, which in essence determines a combined expression level cut off for this set of genes that can predict a new unknown sample to be either matching a control profile or a bacterial infection profile. Leave-One-Out bootstrapping within the study data set determined a (not unexpected, given the training and testing of the classifier happening with the same set of samples) very robust infection-status prediction accuracy of 100%. We further investigate and confirm that for a further 26 independent samples (collected after the original study) from the Royal Infirmary Edinburgh Neonatal Unit the prediction accuracy (that is, both sensitivity or the ability to correctly identify infection, and specificity or the ability to correctly identify controls) for bacterial infection remains at 100%. We also obtain initial evidence that viral infections are not classified as bacterial infections, although at n=3 this determination requires further confirmation.

1.2 Independent validation of classifier

Given that a single study and population is insufficient to devise, validate and further develop a biomolecular classifier towards clinical utility, we pursued additional validation steps. These steps included (a) testing the classifier prediction outcomes on external samples not derived from the original study population and demographic, (b) testing the classifier predictions on an assay platform closer to clinical laboratory use, (c) testing classifier predictions with a reduced number of laboratory steps (measuring gene transcription straight from blood without RNA extraction), and finally (d) testing the classifier in a multi-centre prospective controlled study including other clinical presentations in sepsis (e.g. viral infections).

With regard to further external microarray study based samples, we collated further Edinburgh neonatal samples, samples from an MRC infant pneumonia study in The Gambia (Howie et al, 2014) and an infant sepsis studies in Cincinnati (Wong et al, 2009a; Wong et al, 2009b). Results

noted here are for our gene classifier trained on our Edinburgh neonatal data and this trained classifier then directly tested (training and test samples z-transformed to accommodate different microarray platforms) on the external samples. Our results (manuscript in preparation) suggest that performance of our classifier can vary with the precise study conditions. For example, increasing sensitivity with severity of pneumonia in the Gambia study, better sensitivity in samples taken 1 day post infection compared to 3 days post infection in the Cincinnati study. Both sensitivity and specificity of prediction remain above 90% for Edinburgh neonates, specificity close to 100% in Gambia and Cincinnati samples, and specificity generally above 85% in all studies. We note an exception to this in Cincinnati samples taken 3 days post-infection, with classification sensitivity dropping to 54%. For classifier testing on a platform closer to clinical utility, we abandoned whole-transcriptome microarrays and carried out studies with much smaller QuantiGene Plex Assay (Affymetrix) containing only classifier genes. After establishing new laboratory protocols and progressing the QuantiGene assay design through several versions (currently 48 gene probes), numerical testing is ongoing. Testing and predicting samples through a Leave-One-Out procedure generates predication accuracy >95%, and this is awaiting confirmation in testing on independent samples. We have also successfully used this QuantiGene assay to test host gene expression directly from whole blood neonatal samples (that is, no RNA extraction step), and a manuscript for results from this platform as well as the external validation above is in preparation.

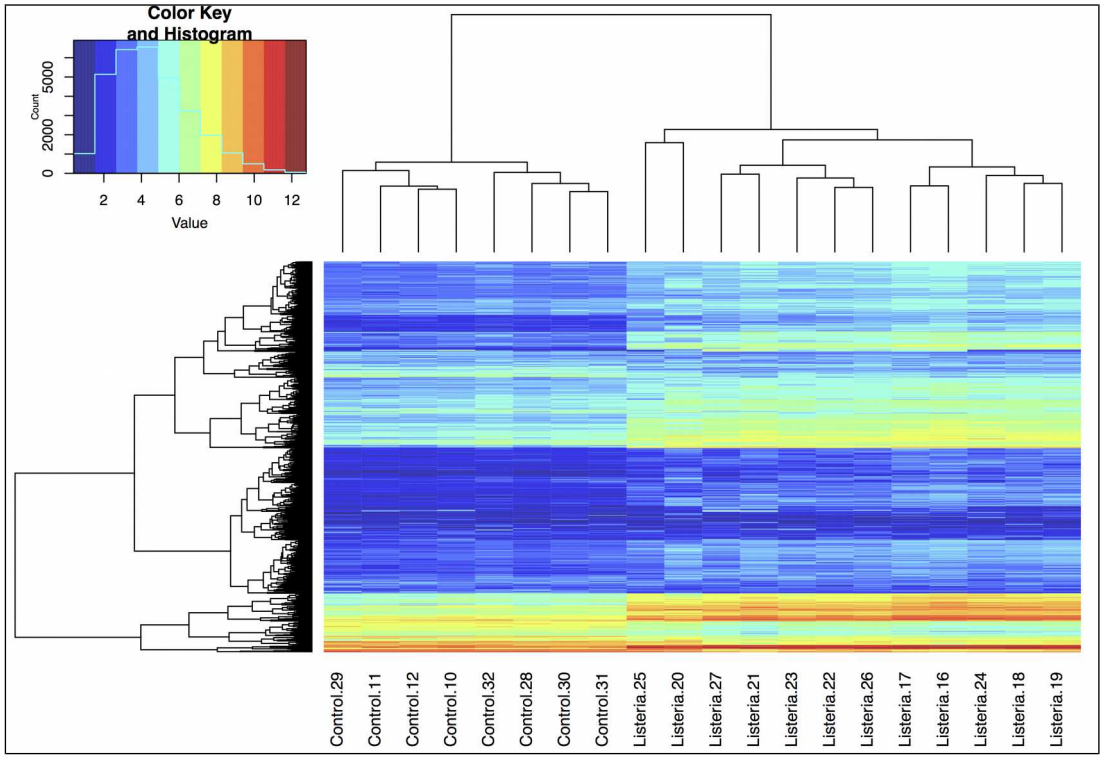
1.3 Sequenced bacterial pathogen genomes

In order to be able to relate a host's immune response to the infecting pathogen, researchers from NS, CIT and DIPM collaborated to sequence the pathogenic strains from a subset of neonatal patients. After confirming the species of clinical isolates from a bacterially infected neonate through microbiological means, isolates were grown and underwent sequencing on an Illumina MiSeq instrument. Subsequent software processing steps of sequence assembly, fragment analysis and automated reported were carried out with the NSilico-developed Simplicity software. This analysis identified each of the six clinical strains as Staphylococcus epidermis, and their draft sequences were made available in public repositories (Kropp et al, 2014a; b; c; d; Kropp et al, 2014e; f). We identified strain to strain mutations that may indicate differences in antimicrobial resistance that would be highly interesting to clinical practice, but in this instance have been unable to match our sequenced pathogen isolates to individual patients and dates, which we elaborate on further below. We have also sequenced the genomes of bacteria in a neonatal mouse model study (detailed below), and generated a bioinformatics pipeline to identify from these particular antimicrobial resistance genes and virulence genes. This however is a proof of principle and involves laboratory strains of the virus used to infect mice. Integration of host and pathogen requires the mapping of a given patient's gene expression profile to the corresponding sequenced pathogen for that patient. Due to ethical and data collection obstacles (detailed below) we are currently unable to perform this mapping, but we are collecting ethical approval and prospective samples for which this can be done.

1.4 Model system for bacterial infection mechanisms

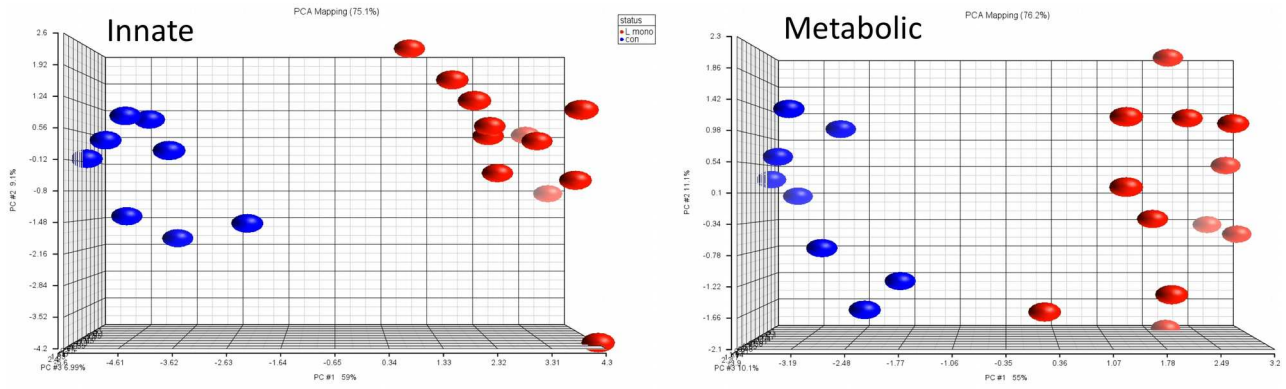
We predicted that part of the presentation of a molecular diagnostic for bacterial sepsis in neonates would include hypotheses or validated hypothesis on the specific biological mechanisms that underpin the neonatal host response to a bacterial infection. To this effect, we collected (from a subset of patients) clinical isolates of bacteria of individual patients and sequenced these isolates in order to identify patient gene expression differences that could be correlated to a particular host gene expression response. However, a particular challenge of ethical consent and clinical laboratory practice became evident (detailed below) in that we were unable to collect sufficient number of

samples and subsequently unable to map patient transcriptome samples to individual patient pathogen sequence data. Recognising this as a persistent issue for future investigations, we directed some of our efforts towards establishing a neonatal mouse model system that could be used in place of human neonates to investigate biological mechanisms of host-pathogen interactions. Based on a collaboration with Prof. Mathias W Hornef (RWTH University Hospital Aachen & Freie Universität Berlin, Germany) we studied the transcriptional response of mouse pups to intranasal bacterial infection with Listeria monocytogenes, comparing infections in 12 neonatal mice (C57BL/6N strain) with 8 neonatal mouse controls. Subsequent to blood collection at three days post-infection and whole-genome microarray analysis (Affymetrix Mouse Gene 2.1 ST Array Plate), we first note that using the most variable genes (standard deviation >0.6 across all samples) separates the data into distinct clusters for control and Listeria-infected samples (Figure 3). Separately, we then identify 46 of our 52 classifier genes on this array platform and cluster – based on a Principal Components analysis - samples based on their gene expression levels (Figure 4). We find that, comparable to human neonatal samples, genes belonging to the innate immune and metabolic response to infection separate the neonatal mouse samples into distinct control and infection clusters.



Rows in this heatmap represent genes, columns represent study samples. Each cell represents the gene expression level of a given gene in a given sample, from low expressed (blues) to high expressed (reds). Dendrograms represent the imposed hierarchical clustering based on similarities between genes and between samples.

Figure 3. Hierarchical clustering of most variable genes



The first two components of a principal components analysis of either innate or metabolic host gene expression in neonatal mice are shown on x and y-axis, respectively. Blue dots are control samples, red dots are Listeria-infected samples.

Figure 4. Principal components analysis of neonatal mouse

Based on the data shown here and additional work on immune cell markers confirming the above, we therefore tentatively propose (manuscript with additional findings submitted to Nature Medicine) that a neonatal mouse model can potentially be used for study of our classifier in experiments for which ethics approval has not been or cannot be obtained easily, where this is primary the case because of patient confidentiality and laboratory data exchange. As a beginning for a further body of work, we have tested a whole genome sequencing analysis on the laboratory strain of Listeria monocytogenes used here, in order to compare antimicrobial resistance genes to data we will be obtaining for human clinical pathogen isolates (one human case of Listeria monocytogenes has to date been identified amongst our study population).

1.5 Algorithms and analysis workflows

Prior to this project but in support of the neonatal sepsis molecular classification techniques, we established a computational requirement for the number of gene predictors to be between 4 and 19 for high accuracy predictions of sample class (Khondoker et al, 2010). For our later neonatal sepsis paper we used those simulation approaches to also establish that the use of a comparatively simple ROC-based classification algorithm (Lauss et al, 2010) compares well to classifiers that are more complex or require separate tuning and optimisation steps (KNN, SVM, Random Forest). Using this classifier therefore alleviates problems in subsequent uses (including retraining the classifier on different gene subsets due to the non-availability of some gene probes on other gene expression platforms) on different data sets or across different microarray and other laboratory data platform, which might otherwise depend on the choice of tuning parameters. To our specification of a classifier algorithm, we also added the requirement that it be robust to sample-wise data standardization (z-score) of samples in order to enable testing a classifier trained on one set of data to be tested on a different data platform. NSilico also developed several reproducible sequencing analysis pipelines that allow researchers to take a sequencing data file (as obtained from a sequencing instrument) to final report (comparative genomics, phylogenetic trees, antimicrobial resistance markers, virulence markers) without having to understand coding or command line tools and with a minimum of parameter inputs. Details of this have been (Walsh et al, 2013) or are in the process of being published (Mullins et al, 2016, submitted to CERC Proceedings). For analysis of host transcriptome data prior to classification algorithm work, existing standards were applied (Forster, 2014; Forster et al, 2003) to statistically test hypotheses of differential gene expression and pathway analysis of gene lists.

1.6 Software solutions and computational infrastructure

For sequencing pathogens and handling data exchange, NSilico Life Science established infrastructure and software to allow non-specialists to carry out analytical workflows (Walsh et al, 2013). The end product "Simplicity" combines command line tools for the reading, assembly and analysis of NGS data (at this time, for prokaryotic organisms), but keeps its complexities hidden from the biological researcher, who is only required to provide the sequencing input file, choose a type of analysis and any basic parameters if required for the analysis. Simplicity is built on a principle of scalability and formally guided by software engineering principles in order to continuously and consistently allow the addition of functionality (different analyses and new bioinformatics tools) and the future increases in data size (moving towards sequencing of eukaryotic genomes and larger studies). To date it has proven invaluable at very quickly turning biological samples from sequence to published manuscripts of draft genomes as well as strain mutation information we can in future use to map against corresponding patients' transcriptional response to infection.

1.7 Integration of host and pathogen biology

We use next generation sequencing on whole-blood derived clinical isolates of bacteria from individual hosts to determine genomic links between host immune response and pathogen. We performed genome sequencing analysis of 6 Staphylococcus epidermis isolates from our population of human neonates. In addition, we have also recently performed a sequence analysis of the laboratory strain of Listeria monocytogenes used to infect hosts in our mouse model study. In order to be able to link host gene responses from a given individual to sequencing data obtained from the bacterial strain infecting the same patient, we developed a bioinformatics analysis pipeline to use resistome and phylogenetic information in the identification of potential antimicrobial resistance genes or biomarkers in the bacterial genome sequence (work submitted to CERC Proceedings 2016, Mullins et al). However, integration of host and genome data are currently limited because we are unable to map individual pathogen genomes to the patient they originate with. We have begun efforts to collect further ethical approval and data sharing capabilities to enable this mapping and therefore the required integration effort for neonatal samples collected now and in the near future.

1.8 Researcher exchange

The expanded ClouDx-i project team consists of investigators and researchers in private industry and academia with backgrounds in biology, medicine, data science, computer science and software engineering. Physical secondments of variable duration took place between NS, CIT and DIPM. These took the form of planned training and knowledge exchange (e.g. software engineering for integrating analysis code into commercial software, transcriptomic analyses), but also included continuous exposure to a different working environment (e.g. academic research department vs private industry research and development) that allows for the casual exchange of knowledge and procedures common within workplaces. This is further elaborated in our conclusions below.

2 Conclusion and future work

The development of a molecular classifier for clinical utility is a multi-stage slow process and associated with a number of challenges that need to be overcome. Given the clinical, biological, analytical, computational and commercial aspects involved, we would suggest that this combination of skills is usually not centralized in one research group and requires collaborative outreach with a common objective albeit different specialisations. We here summarise our conclusions for research

outcomes and challenges in taking this research from biomedical study and biomarker identification towards clinical practice.

Establishment of classifier and biological mechanisms. We have successfully developed a host RNA transcription based classification process that requires (up to) 52 distinct innate immunity, adaptive immunity and metabolic genes and a straightforward ROC based classification algorithm to apply these as a predictor for the host response in an unknown blood sample. Establishing this in a limited study population (Edinburgh neonates) and validating with independent study data (not designed for this purpose) serves as an expanded proof of concept, but in order to satisfy requirements for clinical utility and commercial viability, it is still essential to (a) test the classifier gene set in a large multi-centre UK study, to (b) advance the assay to a purpose-made small and streamlined platform with faster processing times and lower cost, and (c) to illuminate the biological host and pathogen mechanisms that underpin this neonatal host response to bacterial infection. While work on a more advanced assay is in progress, the multi-centre study is currently being prepared as a grant submission. When accepted this will allow us to establish not only the accuracy of our molecular classifier in a more representative UK population sample, but it will also aid the testing of biological sample processing steps, the delineating of host response to viral and other infections (with these not benefitting from antibacterial treatment) rather than bacterial infection, the biological separation between neonatal and adult host responses to bacterial infection, and a more detailed investigation of clinical and demographic variables (age, drugs, weight) in the prediction accuracy.

Development of algorithms and workflows. For clinicians and biologists, selection and implementation of algorithms can be barriers to research progress, as they require a distinctly different domain knowledge, while on the other end service-type analyses (where analysts have no biological or clinical domain knowledge) form a similar barrier. Within ClouDx-i we have addressed this by incorporating data scientists with research interests and knowledge in biomedical application. This both allowed the development and use of algorithm to match objectives in the development of a molecular classifier for diagnosis of neonatal sepsis, and the sharing of knowledge and approaches in the concurrent researcher exchange program. Multiple algorithms and workflows were considered for developing our molecular classifier. We compared four classification algorithms and for all work presented here focused on a ROC-based classifier due to its match to a list of specifications we devised, primarily its relative ease of (repeated and frequent) use and lack of parameters to be tuned on every use. Statistical analysis and machine learning approaches are standardized (Forster et al, 2003) and most importantly were supplemented by a biological pathway analysis to identify not only statistically significant differentially expressed genes, but to target biologically relevant genes in host responses to infection. Most development-intense were analysis pipelines for NGS data, which yet have to approach standardization, but where efficiently and professionally aggregated by NSilico software engineers. Importantly, the specific expertise here is not to develop new ways of processing and analyzing NGS data, but to usefully combine already established software tools by their inputs and outputs in such a way that the execution of a full analysis from original sequencing data file to analysis report is seamless and can be done without any prior knowledge (of course, knowledge of the research question is expected) of command line tools, algorithms and computer systems administration. On availability of a larger and patient-linkable set of pathogen genomes we will research and add to this integration approaches and visualisations for human host gene transcription data and pathogen genome data.

Development of computing infrastructure and software engineering. The understanding of interactions between host and pathogen are a valuable component for understanding biological

mechanisms of host response. Collaboration between the University of Edinburgh and Cork Institute of Technology and NSilico Life Science has been invaluable for the sequencing of clinical isolates from neonate blood samples. Crucially, this technology is more advanced but less standardized than microarray technology (which was used to measure host transcriptional response), it requires computational capacity in term of storage and processing speeds, but also effective software pipelines for daisy-chaining multiple bioinformatics analysis stage tools into one workflow. For the needs of ClouDx-i, NSilico developed Simplicity, which allows lay users without programming knowledge to take sequenced genomes from input file to primary results, and for the purposes of understanding our molecular classifier was used to analyse the genomes of bacterial isolated from a subset of our human neonates and the genome of the laboratory strain used to infect the neonatal mice in our mouse model study. With proof of principle operation and functioning software established, this will now find frequent application in the sequencing of bacterial and viral genomes of individual patients collected currently and prospectively for a multi-centre study.

Exchange of knowledge and skills. It is currently common to use laboratory techniques of high complexity in protocols and analysis in order to determine the sequence of whole genomes and the activity of genes within whole genomes. This requires advanced knowledge of data science to accommodate these technologies. While remote collaboration is useful and necessary, we have usefully promoted the exchange of researchers and computer scientists between institutions for training and knowledge exchange. Within ClouDx-i we have observed mutual benefit. Within this project, a data scientist contributed analytical and statistical expertise to the development of bioinformatics pipelines for next generation sequencing data while at the same time being provided with expertise in processing NGS data and placing analysis solutions within a software engineering framework. A biologist was given the opportunity to carry out laboratory processes for sequencing genomes while contributing basic infection host biology. A second biologist has been providing use cases and biological expertise to the interpretation of NGS data, while receiving instruction in bioinformatics approaches. A software engineer received instruction in the data science aspects of transcriptomic analysis and provided practical guidelines for commercial-standard development of analytical code. On their return to host institutions, these and other individuals were and are adding new capacities. For NSilico, these involve the analysis end of NGS, understanding customer needs and biological interpretation of results. For DIPM, they involve processing of NGS data, use of bioinformatics tools and the ability to contribute program code of the required standard to commercial analysis pipelines. Post-secondment, ongoing interactions involve mutual manuscript review and code exchange, and interactions on research funding applications.

Use of clinical study population. Patient-centric studies are invaluable and of course a pre-condition for our objectives, but they also posed challenges that were not always clear at the outset of the project. Firstly, and of particular sensitivity when studying neonates, we allocated more time and resources than planned to standardizing a low-impact blood sampling and processing procedure, requiring compromise between clinical staff and laboratory needs. Secondly, we noted that patient recruitment rates were affected by changes in clinical diagnosis and treatment protocols, prolonging the time needed to establish a study population of sufficient size and composition. Finally, we noted that the collection of initially unplanned samples became necessary in validation steps of the work, but had not planned for details of data exchange between laboratories, analysts and commercial partners. Specifically, we are to date unable to link host transcriptomic data to pathogen sequencing data. While the latter has been carried out for bacterial isolates from six patients and subsequently published, we were unable to match laboratory identifiers for these samples to clinical patient identifiers. We were also unable to track patient paths within a clinical ward, which would have allowed us to incorporate this information with strain sequence mutations. This is due in part to

ethics approval, where the exact type of linking of data was not predicted at the outset and therefore not included in the ethics application, and in part to non-matching or simply non-collected patient and patient sample identifiers that can logistically and administratively not be disclosed between different clinical laboratories and research institutions. We would therefore recommend for studies of patient-centric development of molecular diagnostic tools that though is given to future sample data collections, data flow between institutions and ethics approval applications that incorporate these.

Future work towards clinical utility. The duration of one EU-funded project has been sufficient to establish a gene set for diagnosing bacterial infections, for establishing the involvement of innate immunity, adaptive immunity and metabolic genes in this host response, and for developing the analysis knowledge and capacity to easily expand this work. However, clinical utility is a more complex process with a higher standard of proof, more technical development, and further knowledge of the biological underpinnings of host-pathogen interactions. While infrastructure and capacities are in place, recruitment of national neonatal care units is now ongoing prior to obtaining funding for a multi-centre UK study for the evaluation of a molecular diagnostic for neonatal bacterial sepsis. Simultaneously, we have recently recruited n>=8 Edinburgh neonates with confirmed (Cytomegalovirus) infection and matched controls in order to train and test molecular classification by distinguishing host immune response to bacterial infection from host immune responses to viral infection. In line with challenges identified above, we will attempt to sequence the genomes of the virus isolated from these patients, but this is subject to separate ethics approval for combining different types of patient data.

Due to the complexity of this research programme, only the combined outcomes of our existing and future research will determine if early communications with diagnostic assay providers proceed to final use in clinical laboratories. We expect that even if commercial expectations are not met in the clinical or cost-effectiveness of a diagnostic assay based on host responses, our efforts will contribute important knowledge to the neonatal host response in bacterial and viral infections, and a fuller understanding for the genomic interactions between hosts and pathogens in patient outcomes.

3 Acknowledgment

We thank the parents and infants involved in these studies, Mathias Hornef, Natalia Torow for providing data in input for the mouse model study. This work was supported by EU FP7-PEOPLE-2012-IAPP (324365), Wellcome Trust (WT066784) programme grant, UK Chief Scientists Office (ETM202) and BBSRC (BB/D019621/1).

4 References

Andrade, S. S., Bispo, P. J. & Gales, A. C. (2008) Advances in the microbiological diagnosis of sepsis. *Shock*, 30 Suppl 1, 41-6.

Cohen, J. (2002) The immunopathogenesis of sepsis. *Nature*, 420(6917), 885-91.

Forster, T. (2014) *Statistical modelling of masked gene regulatory pathway changes across microarray studies of interferon gamma activated macrophages*. PhD The University of Edinburgh.

Forster, T., Roy, D. & Ghazal, P. (2003) Experiments using microarray technology: limitations and standard operating procedures. *Journal of Endocrinology*, 178, 195-204 ST - Experiments using microarray technol.

Howie, S. R., Morris, G. A., Tokarz, R., Ebruke, B. E., Machuka, E. M., Ideh, R. C., Chimah, O., Secka, O., Townend, J., Dione, M., Oluwalana, C., Njie, M., Jallow, M., Hill, P. C., Antonio, M., Greenwood, B., Briese, T., Mulholland, K., Corrah, T., Lipkin, W. I. & Adegbola, R. A. (2014) Etiology of severe childhood pneumonia in the Gambia, West Africa, determined by conventional and molecular microbiological analyses of lung and pleural aspirate samples. *Clin Infect Dis*, 59(5), 682-5.

Hyatt, G., Melamed, R., Park, R., Seguritan, R., Laplace, C., Poirot, L., Zucchelli, S., Obst, R., Matos, M., Venanzi, E., Goldrath, A., Nguyen, L., Luckey, J., Yamagata, T., Herman, A., Jacobs, J., Mathis, D. & Benoist, C. (2006) Gene expression microarrays: glimpses of the immunological genome. *Nat Immunol*, 7(7), 686-91.

Khondoker, M. R., Bachmann, T. T., Mewissen, M., Dickinson, P., Dobrzelecki, B., Campbell, C. J., Mount, A. R., Walton, A. J., Crain, J., Schulze, H., Giraud, G., Ross, A. J., Ciani, I., Ember, S. W., Tlili, C., Terry, J. G., Grant, E., McDonnell, N. & Ghazal, P. (2010) Multi-factorial analysis of class prediction error: estimating optimal number of biomarkers for various classification rules. *J Bioinform Comput Biol*, 8(6), 945-65.

Kingsmore, S. F., Kennedy, N., Halliday, H. L., Van Velkinburgh, J. C., Zhong, S., Gabriel, V., Grant, J., Beavis, W. D., Tchernev, V. T., Perlee, L., Lejnine, S., Grimwade, B., Sorette, M. & Edgar, J. D. (2008) Identification of diagnostic biomarkers for infection in premature neonates. *Mol Cell Proteomics*, 7(10), 1863-75.

Kropp, K. A., Lucid, A., Carroll, J., Belgrudov, V., Walsh, P., Kelly, B., Smith, C., Dickinson, P., O'Driscoll, A., Templeton, K., Ghazal, P. & Sleator, R. D. (2014a) Draft Genome Sequence of a Serratia marcescens Strain Isolated from a Preterm Neonatal Blood Sepsis Patient at the Royal Infirmary, Edinburgh, Scotland, United Kingdom. *Genome Announc*, 2(5).

Kropp, K. A., Lucid, A., Carroll, J., Belgrudov, V., Walsh, P., Kelly, B., Smith, C., Dickinson, P., O'Driscoll, A., Templeton, K., Ghazal, P. & Sleator, R. D. (2014b) Draft Genome Sequence of a Staphylococcus warneri Strain Isolated from a Preterm Neonate Blood Sepsis Patient at the Royal Infirmary, Edinburgh, Scotland. *Genome Announc*, 2(5).

Kropp, K. A., Lucid, A., Carroll, J., Belgrudov, V., Walsh, P., Kelly, B., Smith, C., Dickinson, P., O'Driscoll, A., Templeton, K., Ghazal, P. & Sleator, R. D. (2014c) Draft Genome Sequence of a Streptococcus agalactiae Strain Isolated from a Preterm Neonate Blood Sepsis Patient at the Royal Infirmary, Edinburgh, Scotland. Genome Announc, 2(5).

Kropp, K. A., Lucid, A., Carroll, J., Belgrudov, V., Walsh, P., Kelly, B., Smith, C., Dickinson, P., O'Driscoll, A., Templeton, K., Ghazal, P. & Sleator, R. D. (2014d) Draft Genome Sequence of an Enterococcus faecalis Strain Isolated from a Neonatal Blood Sepsis Patient. *Genome Announc*, 2(5).

Kropp, K. A., Lucid, A., Carroll, J., Belgrudov, V., Walsh, P., Kelly, B., Templeton, K., Smith, C., Dickinson, P., O'Driscoll, A., Ghazal, P. & Sleator, R. D. (2014e) Draft Genome Sequence of a Pantoea sp. Isolated from a Preterm Neonatal Blood Sepsis Patient. *Genome Announc*, 2(5).

Kropp, K. A., Lucid, A., Carroll, J., Belgrudov, V., Walsh, P., Kelly, B., Templeton, K., Smith, C., Dickinson, P., O'Driscoll, A., Ghazal, P. & Sleator, R. D. (2014f) Draft Genome Sequence of a Staphylococcus aureus Isolate Taken from the Blood of a Preterm Neonatal Blood Sepsis Patient. *Genome Announc*, 2(5).

Lauss, M., Frigyesi, A., Ryden, T. & Höglund, M. (2010) Robust assignment of cancer subtypes from expression data using a uni-variate gene expression average as classifier. *BMC Cancer*, 10, 532.

Liu, L., Johnson, H. L., Cousens, S., Perin, J., Scott, S., Lawn, J. E., Rudan, I., Campbell, H., Cibulskis, R., Li, M., Mathers, C., Black, R. E. & UNICEF, C. H. E. R. G. o. W. a. (2012) Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet*, 379(9832), 2151-61.

Manger, I. D. & Relman, D. A. (2000) How the host 'sees' pathogens: global gene expression responses to infection. *Curr Opin Immunol*, 12(2), 215-8.

Smith, C. L., Dickinson, P., Forster, T., Craigon, M., Ross, A., Khondoker, M. R., France, R., Ivens, A., Lynn, D. J., Orme, J., Jackson, A., Lacaze, P., Flanagan, K. L., Stenson, B. J. & Ghazal, P. (2014) Identification of a human neonatal immune-metabolic network associated with bacterial infection. *Nat Commun*, 5, 4649.

Stoll, B. J., Hansen, N. I., Adams-Chapman, I., Fanaroff, A. A., Hintz, S. R., Vohr, B., Higgins, R. D. & Network, N. I. o. C. H. a. H. D. N. R. (2004) Neurodevelopmental and growth impairment among extremely low-birth-weight infants with neonatal infection. *JAMA*, 292(19), 2357-65.

Walsh, P., Carroll, J. & Sleator, R. D. (2013) Accelerating in silico research with workflows: a lesson in Simplicity. *Comput Biol Med*, 43(12), 2028-35.

Wong, H. R. (2013) Genome-wide expression profiling in pediatric septic shock. *Pediatr Res*, 73(4 Pt 2), 564-9.

Wong, H. R., Cvijanovich, N., Allen, G. L., Lin, R., Anas, N., Meyer, K., Freishtat, R. J., Monaco, M., Odoms, K., Sakthivel, B., Shanley, T. P. & Investigators, G. o. P. S. S. S. (2009a) Genomic expression profiling across the pediatric systemic inflammatory response syndrome, sepsis, and septic shock spectrum. *Crit Care Med*, 37(5), 1558-66.

Wong, H. R., Cvijanovich, N., Lin, R., Allen, G. L., Thomas, N. J., Willson, D. F., Freishtat, R. J., Anas, N., Meyer, K., Checchia, P. A., Monaco, M., Odom, K. & Shanley, T. P. (2009b) Identification of pediatric septic shock subclasses based on genome-wide expression profiling. *BMC Med*, 7, 34.

Metamorphosis: Changing the shape of genomic data using Kafka

B Lawlor, P Walsh
Department of Computer Science
CIT – Cork Institute of Technology, Ireland
e-mail: brendan.lawlor@mycit.ie

Keywords: Bioinformatics, Big Data, Cloud Computing

A bioinformatic researcher, when seeking novel ways to process genomic data, will naturally look for a standarad reference genomic database to test their new ideas against. Perhaps the best-known genomic reference is NCBI's Blast[1]. The word 'Blast' can refer both to the database and to the search algorithm used to access it, and in fact this ambiguity highlights an underlying problem for the aforementioned bioinformatic researcher. The shape of the Blast genomic data is set by the Blast search algorithm. To make this data available in its basic form – for example to researchers who wish to test other search algorithms or process the entire database contents - it must be downloaded, processed and reformatted into flat FASTA files. In either form – algorithm-specific proprietory format or more than 800GB of text files - the genomic reference data is not in good shape.

Genomic data is big data. It has high volume and increasingly high velocity (more database entries, more query sequences). The format in which that data is maintained matters. It should be robust, scalable and amenable to parallelized access. It should also be available in a way that stimulates innovation. If it is only available through BLAST, then it will only be used in ways that BLAST permits. If instead it is open to exhaustive, streamed, distributed and parallelized access, then novel applications of the data will result.

This paper explores the use of Apache Kafka[2] - a technology that was not conceived as a database but as a messaging system - for general purpose, efficient, distributed, stream-ready storage and retrieval of genomic data. It offers scalable, durable, distributed storage, with high-speed and flexible access. We consider how this technology could emerge as a primary format for storing genomic sequences, directly accessible by general purpose programs, but from which secondary algorithm-specific formats could be derived where needed. We present a Proof of Concept implementation of a Kafka Genomic Database and measure its properties.

The wider software community is exploring improved ways of storing big data that avoid the bottlenecks of centralized transactional databases. A central feature of such systems is the re-interpretation of system state as 'events'. 'Event sourcing'[3] is an increasingly popular and scaleable solution to maintaining the state of a system without creating database bottlenecks[4]. Reference genomes can be seen as the system state of many bioinformatics applications, and a Kafka-based log of the adds, updates and deletes of genomic sequences could offer a powerful implementation of this concept.

It is interesting to note that Event Sourcing as described above is analogous to transaction logs – the history of actions maintained by traditional database management systems to guarantee ACID properties[5].

References

[1] Altschul, Stephen F., et al. "Basic local alignment search tool." Journal of molecular biology 215.3 (1990): 403-410.

[2] Wang, Guozhang, et al. "Building a replicated logging system with Apache Kafka." Proceedings of the VLDB Endowment 8.12 (2015): 1654-1655.

[3] Martin Fowler. Event sourcing. http://martinfowler.com/eaaDev/EventSourcing.html, 2005.

[4] Betts, Dominic, et al. "Exploring CQRS and Event Sourcing: A journey into high scalability, availability, and maintainability with Windows Azure." (2013).

[5] Gray, Jim, and Andreas Reuter. Transaction processing: concepts and techniques. Elsevier, 1992.

A Generalized Multi-Network Framework for Disease Gene Progression Prioritization

Fiona Browne [1], Haiying Wang [1], Huiru Zheng [1], Juan Carlos Leyva López ac [2], Jesús Jaime Solano Noriega [2]
[1] School of Computing and Mathematics
Ulster University, Northern Ireland
[2] Universidad de Occidente
Blvd. Lola Beltrán y Blvd. Rolando Arjona, Culiacán, Sinaloa, México
e-mail: f.browne@ulster.ac.uk

Keywords: outranking methods, multi-objective evolutionary algorithms, multi-network analysis

Background

Identification of disease progression and disease genes associated with physiological disorders is a fundamental task in the analysis of complex diseases. Diseases are often polygenic and multifactorial in nature and can present with different genetic perturbations in patients [[1]]. Recent years have witnessed the systematic investigation of complex disease through the application of high-throughput experimental technologies and the development of centralized databases. These experimental "omic" platforms target the comprehensive analysis of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics). This research has been essential in (i) advancing the knowledge of biological systems (2) understanding, predicting, diagnosing and monitoring diseases (3) discovering biomarkers and (4) identifying drug targets. However, translation of relevant prognostic markers identified by such experiments into clinical tools for personalized patient treatment has been slow due to issues such as the reproducibility and validity of findings across studies, unfocused study design and inappropriate selection and application of statistical techniques (Dupuy & Simon 2007) (Simon 2005). In addition, experimental validation of large lists of candidate genes lists returned from experimental techniques are time-consuming and expensive with data obtained from individual platforms insufficient to fully characterize complex biological systems. The challenge we now face is the development of efficient methodologies to prioritize these gene lists and address the limitations of single platforms. To this end, combining experimental results from multiple "omic" platforms can help to uncover latent biological relationships evident only through integration of measurements across multiple biochemical domains (Wanichthanarak et al. 2015) (Hoadley et al. 2014). Network analysis has proved to be a very effective approach in modelling complex connections among diverse types of cellular components such as genes and metabolites.

[1] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4686785/#B1

Proposed Framework

In this research, we present a generalized framework (Figure 1), which integrates diverse "omic" data such as mRNA, gene co-expression, Gene Ontology (GO) semantic similarity and tissue-specific data using a consensus network combination approach. A Case Study based on Ovarian Cancer (OV) disease progression is presented to demonstrate the application of the framework. OV is associated with substantial mortality and morbidity, this disease is difficult to detect at its earliest stage, therefore, uncovering gene signatures at an early stage is an essential task. The framework as illustrated in Figure 1 consists of a number of stages. (A) extracting RNA-Seq OV gene expression and clinical information, (B) identification of differentially expressed genes (DEGs) between early stage and late stage tumour samples, (C) construction of DEG septic networks, (D) analysis of the DEG specific networks to identify stage associated hub genes and their interactors, (E) novel outranking-based algorithm. The method is based on ELECTRE III (Roy 1991) to create a fuzzy outranking relation and then, it uses a multi-objective evolutionary algorithm (López et al. 2016) to exploit the outranking relation and to derive a prioritized list of candidate disease genes. (F) evaluation of hub genes including the development of a hub gene-based classier model to distinguish OV stages.

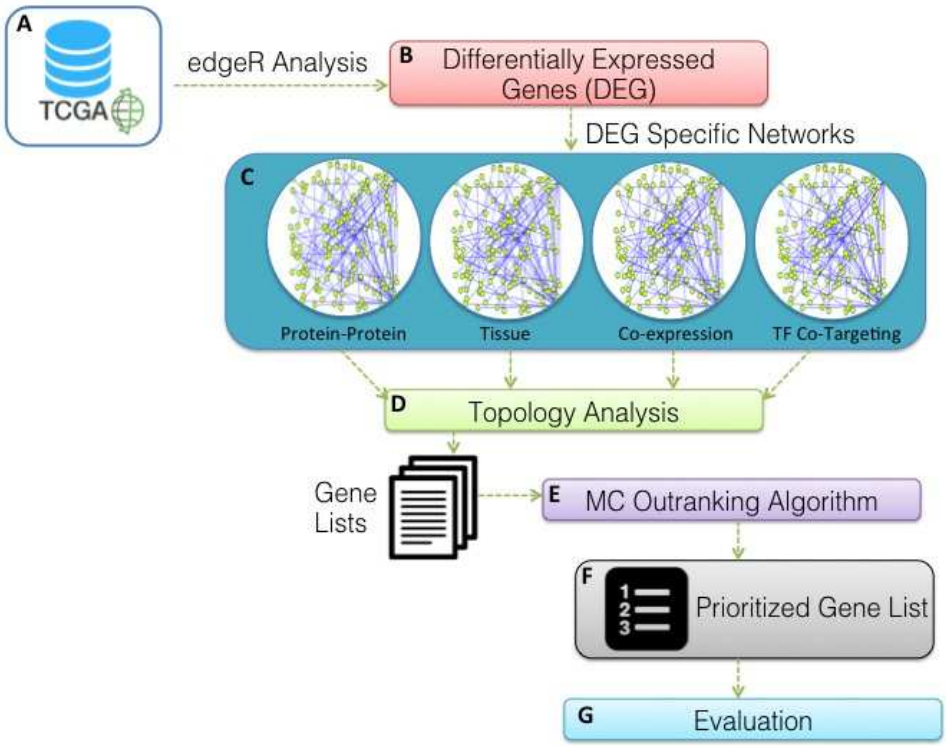


Figure 1, Overview of the Generalised Framework For Disease Progression Gene Prioritization

Preliminary Results

The proposed framework integrated data from four networks: protein-protein, co-expression, tissue and transcription factor co-targeting. Topological analysis based on network degree and betweenness was performed for the individual networks. Lists of genes with high degree and betweenness were selected from each network and then ranked using the novel outranking-based algorithm. A total of 11 genes were identified using the ranking approach. Using the Online Mendelian Inheritance in Man (OMIM) Morbid Map dataset, which lists diseases and gene associations from published studies. A total of 35 known ovarian genes were identified of which 2 genes from the ranked list namely Erbb2 and BRAC1 both of which have been implicated in the carcinogenesis and prognosis of ovarian cancer. Gene Ontology (GO) and pathway analysis was applied to investigate the biological implications of the integrated prioritized disease gene candidate list. Using the GO Biological Process and the DAVID system (Huang et al. 2008) a statistical over representation test was performed using the prioritized genes ($p < 0:05$ Bonferroni corrected), 79 GO processes were identified including ovulation cycle and ovulation cycle process. Furthermore, 16 significant KEGG pathways were identified including endometrial cancer, MAPK signaling pathway, and pathways in cancer. These initial results demonstrate the biological significance of a network based approach in integrating “omic” data.

Conclusions

Network analysis is an effective approach in modeling the complexity of human disease. Furthermore, networks have been instrumental in uncovering how complexity controls disease manifestations, prognosis, and therapy. Our initial studies have illustrated how applying a network approach, diverse “omic” data can be integrated and provides encouraging results for biologically important tasks such as ovarian cancer disease gene prioritization.

References

Dupuy, A. & Simon, R.M., 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. Journal of the National Cancer Institute, 99(2), pp.147–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17227998 .

Hoadley, K.A. et al., 2014. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell, 158(4), pp.929–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25109877.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols, 4(1), pp.44–57. Available at: http://www.nature.com/doifinder/10.1038/nprot.2008.211 .

López, J.C.L. et al., 2016. Exploitation of a Medium-Sized Fuzzy Outranking Relation Based on Multi-objective Evolutionary Algorithms to Derive a Ranking. http://dx.doi.org/10.1080/18756891.2016.1204122.

Roy, B., 1991. The outranking approach and the foundations of electre methods. Theory and Decision, 31(1), pp.49–73. Available at: http://link.springer.com/10.1007/BF00134132

Simon, R., 2005. Roadmap for Developing and Validating Therapeutically Relevant Genomic Classifiers. Journal of Clinical Oncology, 23(29), pp.7332–7341. Available at: http://www.jco.org/cgi/doi/10.1200/JCO.2005.02.8712

Wanichthanarak, K., Fahrmann, J.F. & Grapov, D., 2015. Genomic, Proteomic, and Metabolomic Data Integration Strategies. Biomarker insights, 10(Suppl 4), pp.1–6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4562606&tool=pmcentrez&rendertype=abstract

Measuring the Matches: Genome Alignment and Benchmarking

Xiangwu Lu [1,3], Audrey M. Michel [2], Pavel V. Baranov [2] and Paul Walsh [3]

[1] Department of Computer Science, Cork Institute of Technology, Cork, Ireland

[2] School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland

[3] NSilico Life Science, Bishopstown, Cork, Ireland

xiang.lu@gmail.com, paul.walsh@nsilico.com

Keywords

Sequence alignment, Short read alignment tools, ribosome profiling, ribo-seq, mRNA, ribosome protected RNA, Genome.

Abstract

Sequence alignment is a key bioinformatics technique for identifying DNA, RNA and protein sequences that are similar due to functional, structural or evolutionary relationships. This paper focuses on the identification and analysis of messenger RNA (mRNA), type of RNA molecule, that conveys genetic information from DNA to molecular machinery, known as ribosome. The main function of the ribosome is to link the amino acids of the protein sequence in an order, specified by mRNA. Accurate determination of mRNA molecule is a key step for understanding a wide range of cellular processes. However, the identification of ribosome protected mRNA fragments, which are typically 30 base pairs long (ribo-seq), is difficult due to the presence of the ribosome. The main issue with the analysis of these short ribosome protected RNA fragments, is the shortage of relevant bioinformatics tools, since most of the tools are developed to identify long nucleotide sequences. Therefore, the choice of the tool, for detecting much shorter ribosome protected RNA, is haphazard. This study provides comprehensive analyses of 13 alignment tools for short sequences, and serves as a useful guide on the best techniques for short read based research. To examine and compare alignment tools, we considered the simulated ribo-seq data of a human genome. For the assessment of accuracy and throughput of the tools, genome alignment benchmarking was divided into intra-exon and splice junction mapping, and presented separately. Intra-exon alignment benchmarking showed that bowtie, bowtie2, gsnap, rum and tophat accurately mapped 99.9% of the reads. bowtie (-v mode) had the highest combined accuracy and throughput with 0 or 1 mismatches allowed, achieving the top True Positives (TPs) with 0% False Positives (FPs) and False Negatives (FNs). bwa proved the correspondence between the number of introduced mismatches (IMs) and the allowed mismatches, achieving the highest combined accuracy and throughput with INDEL reads. INDEL reads were most accurately mapped with smalt. However, the alignment of ribo-seq data to an eukaryotic genome, with alternative splicing, was more challenging. In particular, the alignment accuracy assessment of short, 30 bp sequences, that span splice junctions, was difficult because the mapping function for splice junctions was absent from many tools. As only 4 tools, gsnap, rum, star and tophat, had mentioned splice junction mapping capability in their documentation, splice junction benchmarking was carried out using these tools only. tophat and gsnap had the highest accuracy and throughput respectively. tophat (with bowtie2) had the best combined accuracy and throughput, highest TPs and lowest FPs, with supplied genome annotation. All splice junction alignment tools (gsnap, rum, star and tophat) had lower throughput when genome annotation was used for the alignment. In terms of scalability, many tools shared similar throughput between 1 to 10 million reads for both, intra-exon and splice junction alignment. bfast, bwa, soap2 and gsnap had the highest throughput. The results demonstrated that all the tools gained higher accuracy by using the genome annotation file, but the throughput was decreased significantly. Overall, this study provides a comprehensive overview of the appropriate tools for short read alignment using reference genome.

1 Introduction

Ribosome profiling, often referred as ribo-seq (Ingolia et al. 2009), is a technique that provides Genome Wide Information on Protein Synthesis (GWIPS) in vivo (Weiss & Atkins 2011). Ribo-seq deals with sequencing specific mRNA fragments from lysed cells to determine the translated mRNAs (Michel & Baranov 2013; Ingolia 2014). Typically, the ribosome protects approximately 30 nucleotides from ribonuclease digestion (Steitz 1969), although both shorter and longer ribosome protected fragments have been reported by Lareau et al. 2014; O'Connor et al. 2013 andGuydosh & Green 2014. Nevertheless, the majority of the sequences, produced using the above

technique, are constrained by the length of the ribosome mRNA channel to ~30 bp single-end reads. At present, RNA-seq alignment tools are used widely in ribo-seq area, to align short single-end reads against a transcriptome or a genome (Wang et al. 2009). Longer reads usually cover one or more genomic regions in mRNA to reduce the gene isoforms, and are used to assemble a genome or transcriptome. In contrast, shorter read length poses a number of mapping challenges (Dobin et al. 2013; Marco-sola et al. 2012). A short read could be mapped to multiple locations due to the sequence similarity; the ability of mapping short 30 bp sequences across splice junction. The accuracy and throughput of 13 short read alignment tools, listed below, were evaluated by aligning simulated ribo-seq data to a human genome, using several short read alignment tool benchmarking tests (Hatem et al. 2013; Ruffalo et al. 2011; Holtgrewe et al. 2011; Schbath et al. 2012).

The tools examined in this research included: bfast (Homer et al. 2009), blat (Kent 2002), bowtie (Langmead 2010), bowtie2 (Langmead & Salzberg 2012), bwa (Li & Durbin 2009), gsnap (Wu & Nacu 2010), maq (Li et al. 2008), rum (Grant et al. 2011), smalt (Ponsting & Ning 2010), ssaha2 (Ning et al. 2001), star (Dobin et al. 2013), soap2 (Li et al. 2009) and tophat (Trapnell et al. 2009).

1.1 Test condition

Platform. The testing was carried out on a server machine with Intel® Xeon® CPU E5645, 24 processors and 48.3GB RAM; Ubuntu 12.04.2 LTS, 64 bit.

Genome indexes. In order to map the reads to a genome, indexes need to be built by using the reference genome sequence. Most of the candidate tools have the functionality to build indexes when reference genome sequence is supplied by the user. However, two tools represented the exceptions. Blat builds indexes on the fly in the beginning of alignment and tophat uses either bowtie or bowtie2 indexes according to the alignment mode,. Hence, pre-built genomic indexes available from the corresponding websites were used by the following tools: bowtie, bowtie2, gsnap, rum and tophat.

Input parameters. The default settings were applied for the most parameters. However, a few parameters set to appropriate values based on the manual, included: mismatch allowed (bfast: no mismatch option; blat: 0, 1; bowtie: 0, 1, 2, 3; bowtie2: 0, 1; bwa: 0, 1, 2, 3; gsnap: 0, 1, 2, 3; maq: 1, 2, 3; smalt: no mismatch option; soap2: 0, 1, 2; ssaha2: no mismatch option; star: 0, 1, 2, 3; tophat: 0, 1), multi-process (set to 4), alignment report mode (report up to 100 alignments) and seed length (set to 25). We also set the relevant input parameters for the alignment of ribo-seq data (the same parameters were used for the mapping of simulated single-end 30 bp intra-exon and splice junction reads)

Test data. To record the accurate genomic coordinates for the simulated data, 30 bp nucleotide sequences for both intra-exon and splice junction reads were sampled from UCSC DAS server hg 19 databases (http://genome.ucsc.edu/cgi-bin/das/). As shown in Figure 1, the sequences were taken from the exon regions only, the intron and exon annotation were obtained from NCBI refGene annotation (UCSC RefGene.txt.gz, download date: November 02, 2014, (Anon n.d.)). The human genome sequence file (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.2bit, version March 8, 2009) was downloaded from UCSC (Anon n.d.) and converted to FASTA format using the ‘twoBitToFa’ utility (Anon n.d.). This FASTA file was used as input for building genome indexes for some tools.

In order to capture the ability of the tools to handle the single nucleotide polymorphisms (SNPs) or sequencing errors, we also introduced 1 to 3 simulated mismatches (IMs), 1 bp insertion (IN) and 1

bp deletion (DEL) into the 30 bp simulated reads. Consequently, that made up 6 different datasets for the benchmarking. The original randomly sampled sequence did not contain simulated IMs. Each sequence file had FASTA or FASTQ format with high quality sequencing value (I). Since identical quality values were applied to all sequence bases and included SNPs, IN and DEL, the ability of some tools to assess and report the alignment mapping score was not addressed in this work. Two datasets of 1 and 10 million reads were generated to test the scalability of each alignment tool.

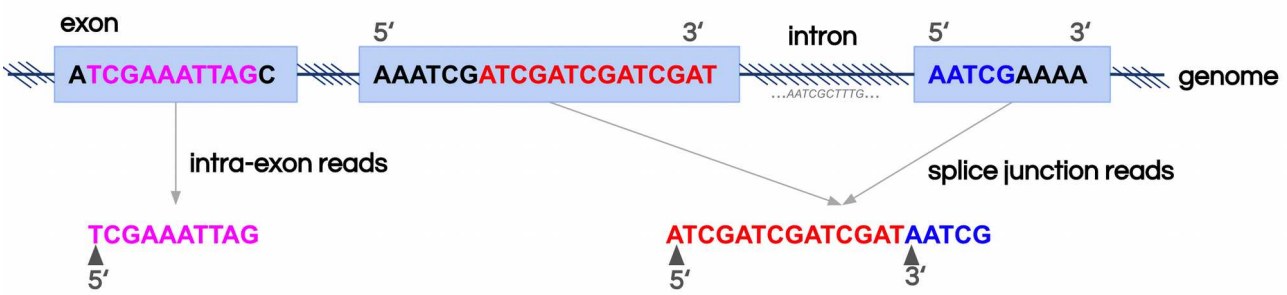


Figure 1. Illustration of sampling principle of intra-exon and splice junction reads from exon regions only. The sequences of a splice junction read in red and blue indicate the 5' and 3' ends respectively. The coordinate of intra-exon 5' location, 5' and 3' of splice junction locations are recorded for evaluation.

Benchmarking procedure. The procedures of intra-exon and splice junction alignment benchmarking workflow are designed separately (Figure 2). In-house Python code was developed to manipulate the workflows and capture the results. By processing the alignment command, the python program splits the alignments into two main categories: unique mappers and non-unique mappers. Important information was derived from the alignment for the verification and classification of several criteria. The alignment showed whether the query read aligned to the correct chromosome and genomic location by checking the 5' end coordinate; or if the sequence distance between query read and reference sequence was less than or equal to the number of allowed mismatch. The alignment was classified into an according group based on the criteria above. Because each splice junction read was sampled from two continuous exons, 3' end of the first exon sequence and 5' end of the second exon sequence (Figure 1), an extra validation condition was added to splice junction alignment benchmarking. This condition included division of a splice junction read into two sections and checking the start coordinates of both section.

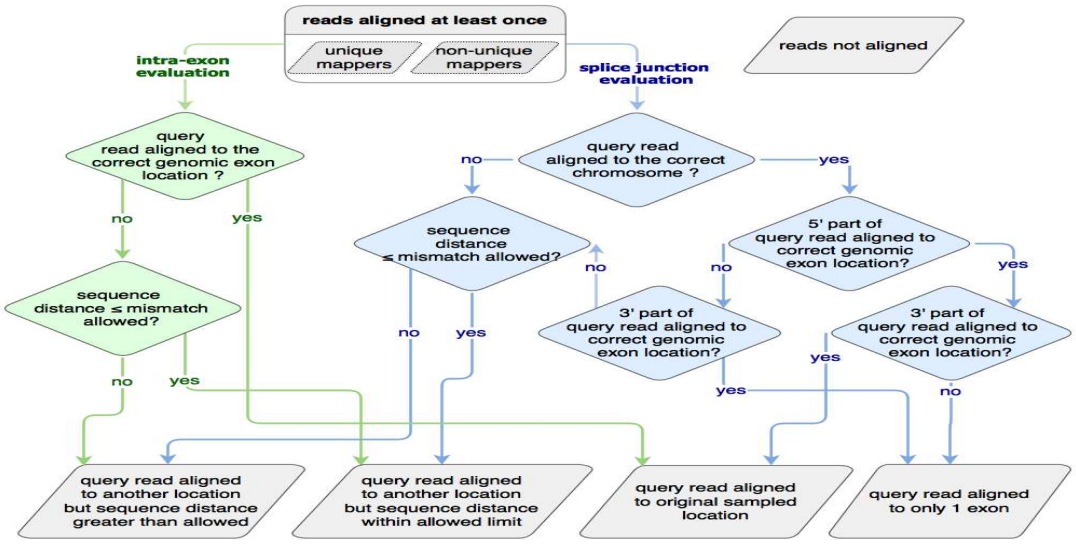


Figure 2. The benchmarking procedure of intra-exon alignment and splice junction alignment evaluation. Alignment(s) of each read were evaluated and the reads were categories in number of different groups.

All the tools were tested for intra-exon alignment benchmarking, but only 4 were selected for splice junction benchmarking, because these tools were reported to be able to align splice junction reads. The results were recorded for performance with and without annotation file (genes.gtf file, version March 18, 2013 downloaded from Illumina iGenomes).

Alignment accuracy. Through our evaluation processes of intra-exon alignment, the query reads had been categories into the following groups: “reads aligned correctly” (query reads aligned to their original sampled location); “reads not aligned” (reads with no alignment); reads aligned to incorrect location” (query reads aligned to locations other than the original sampled location, but the sequence distance between the query reads and the reference sequence was greater than the mismatch allowed). For the splice junction alignment, only if both 5’ and 3’ part of the query reads aligned to their original sampled location were considered as “reads aligned correctly”. We calculated the proportion of three criteria above, by adding the unique and non-unique mappers together.

Alignment throughput. The CPU time (‘read’ and ‘sys’ values from time command) and throughput were recorded separately for intra-exon and splice junction alignments. As aligning to the human genome can take a considerable time, we decided to record the time required to align 2 datasets of different size, one with 1 million and another with 10 million reads, rather than carrying out multiple iterations.

2 Conclusion

2.1 Results of genome intra-exon alignment

Genome intra-exon alignment accuracy. Unlike the transcriptome benchmarking, none of the tools mapped 100% of the simulated ribo-seq reads to their original sampled location when there were 0 IMs in the reads and 0 mismatches allowed (Figure 3). Many of the tools, bowtie, bowtie2, gsnap, rum and tophat, however, did have 99.9% accuracy under these conditions.. As with the

transcriptome results, for some tools there was a strong correspondence between the number of IMs and the mismatch constraint. Bwa provides a good example of this, 95% of the simulated query reads aligned to the original sampled locations when there were 0 IMs and 0 mismatches permitted. However, no query reads aligned when there was 1 IM in the reads and 0 mismatches permitted. After comparing these results to those of soap2, it was clear that ~ 94% of reads were accurately mapped to their original sampled location, when 1 IM was present in the reads but the mismatch allowed parameter was set to 0. With respect to the mapping of reads with INDELs, smalt, bwa, star and tophat, had the highest mapping accuracy in that order (Figure 3).

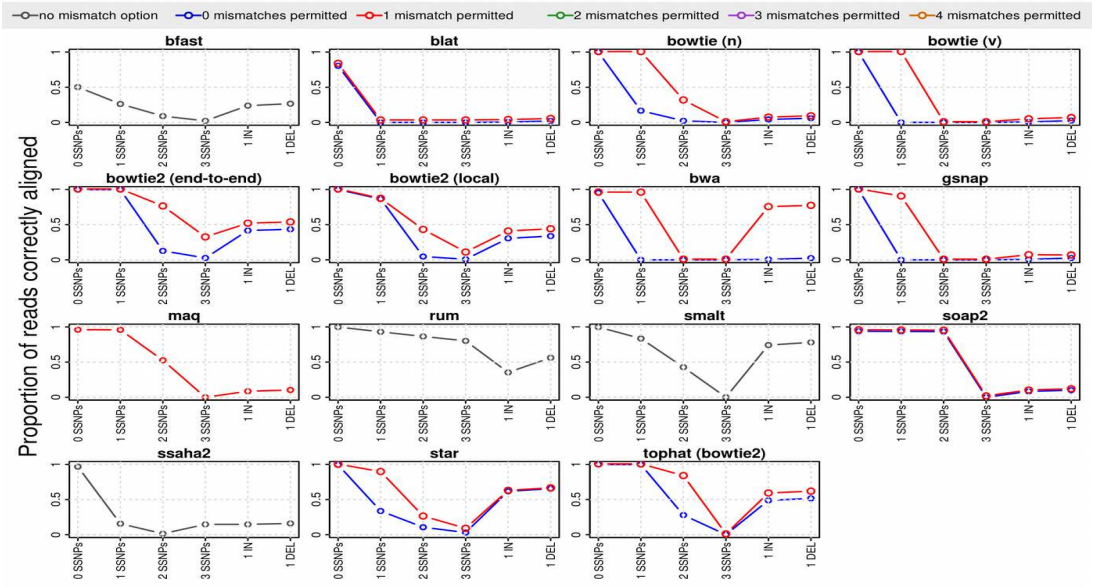


Figure 3. The accuracy of each short read alignment tool evaluated for genome exon alignment using different IM/INDEL and mismatch metrics.

Genome intra-exon alignment throughout. Another aspect we wished to investigate, was the time required to align different size datasets, or the scalability of the tools. Many of the tools had a similar throughput when mapping a dataset of 1 million reads compared to 10 million reads, which indicates that the tools scaled up reasonably well (Figure 4). Also, the majority of the tools did not exhibit great differences in throughput, as the number of IMs/INDELs and mismatches allowed were varied in each case. Purely in terms of alignment throughput, bfast, bwa and soap2 had the highest, although their performance needed to be examined with respect to accuracy (see section Genome intra-exon alignment accuracy and throughput). Note that, while we recorded the throughput of each tool, the memory footprint was not recorded because a tool may achieve higher throughputs at the expense of higher memory usage. In practice, the amount of memory space used by a tool, particularly for large genomes, may be an important consideration.

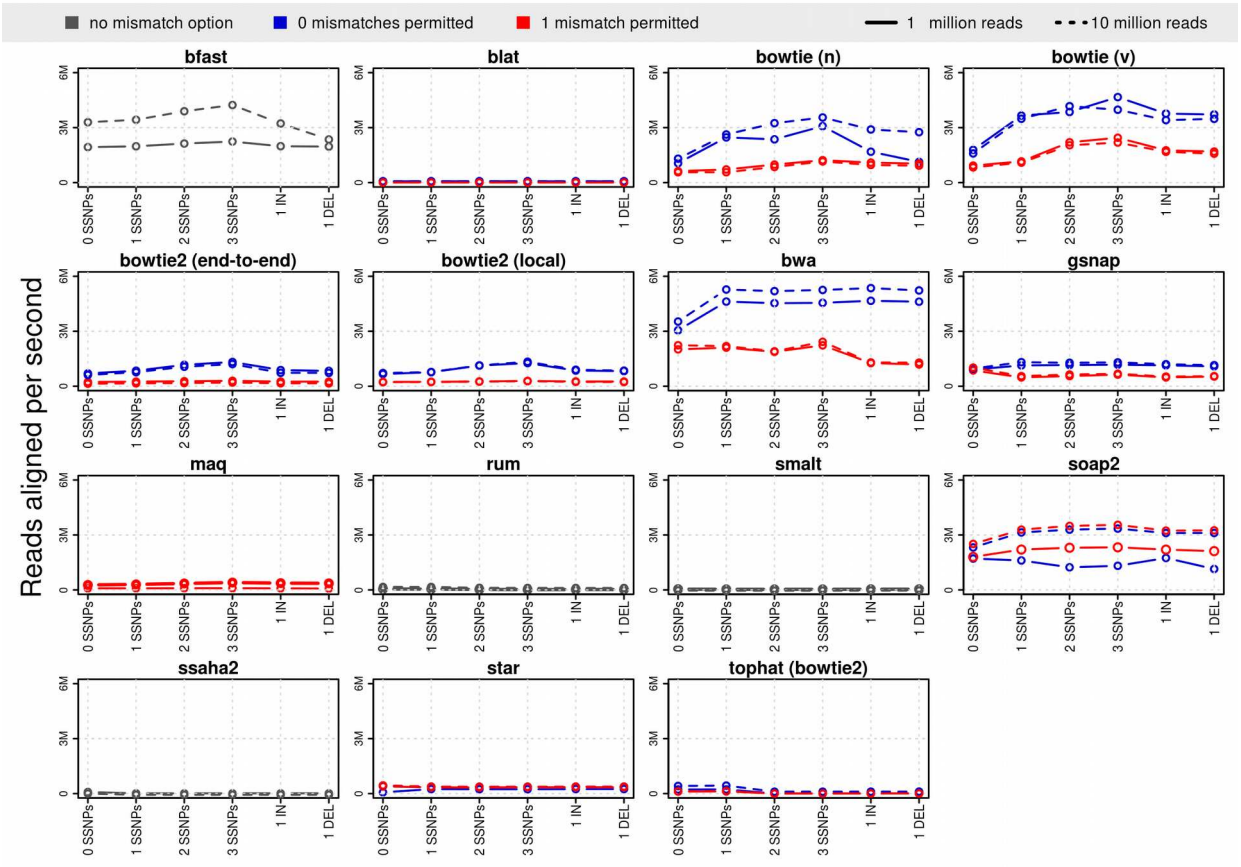


Figure 4. The comparison of the genome alignment throughput of each short read alignment tool. The throughput of mapping 1 million and 10 million reads datasets was compared and evaluated for genome exon alignment using different IM/INDEL and mismatch metrics. "M" in the Y-axis indicates million.

Genome intra-exon alignment accuracy and throughput. Overall, bowtie (-v mode) had the highest combined accuracy and throughput for alignments of simulated ribo-seq reads to the intra-exon only regions of the human genome, when 0 or 1 mismatches were allowed (Figure 5). However, when the INDELs were introduced into the reads, bwa achieved the highest accuracy and throughput.

Figure 5. The accuracy and throughput of each short read alignment tool evaluated for genome exon alignment using different IM/INDEL and mismatch metrics. The size of the pie chart is proportional to the throughput achieved by the alignment tool, while the green slicing of the pie chart represents the alignment accuracy.

Genome intra-exon alignments sensitivity and specificity. The number of True Positives (TPs), False Positives (FPs) and False Negatives (FNs) for each tool for the selected genome intra-exon benchmarking are shown in Figure 6. Under 0 IM and 0 mismatches conditions, bowtie (-v mode), bwa, gsnap, rum and tophat all had very high true positive rates with 0% FPs and 0% FNs, with bowtie (-v mode) topping the list in terms of TPs.

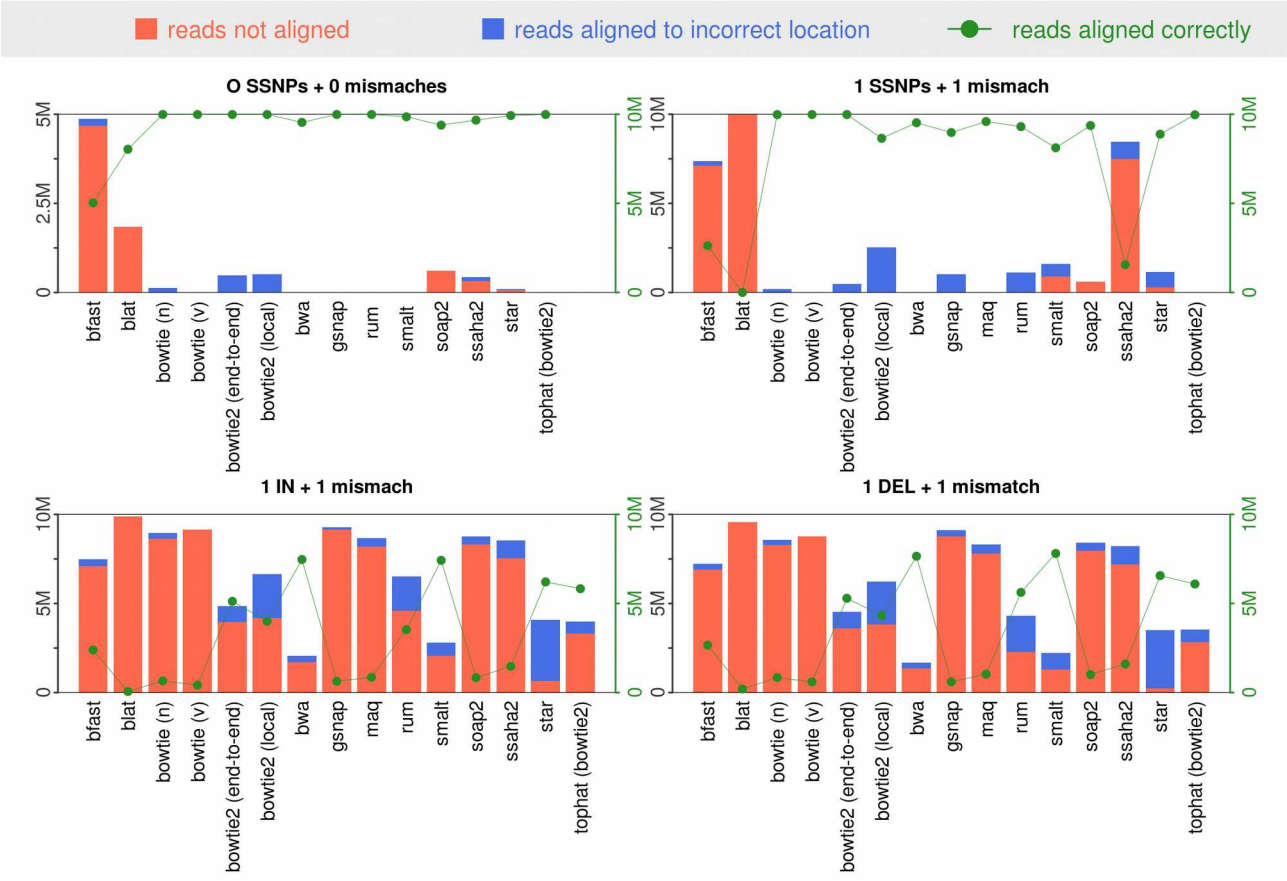


Figure 6.The true positives (TPs), false positives (FPs) and false negatives (FNs) of each short read alignment tool evaluated for genome exon alignment, using different IM/INDEL and mismatch metrics. The green lines represent the number of reads out of 10 million, which were mapped to the original sampled location (TPs). The orange bars represent the number of reads not aligned (FNs) and the blue bars represent the number of reads aligned to another location, where the sequence distance was greater than the mismatch allowed metric. Note that, in the plot of "0 IMs + 0 mismatches", the scale of left Y- axis for the bars is from 0 to 5 million and the remaining Y-axis scales are from 0 to 10 million.

2.2 Results of genome splice junction alignment with simulated ribo-seq data

Genome splice junction alignment accuracy. Figure 7 shows that all 4 short read alignment tools performed much better when transcript annotations were provided for the analysis. It appears that, tophat (with bowtie2) accurately mapped a higher proportion of simulated ribo-seq reads across the known sampled splice junctions, when there were 0 or 1 IMs (Figure 7). Rum also performed reasonably well even with up to 3 IMs in the reads, while it could not handle the INDELs very well.

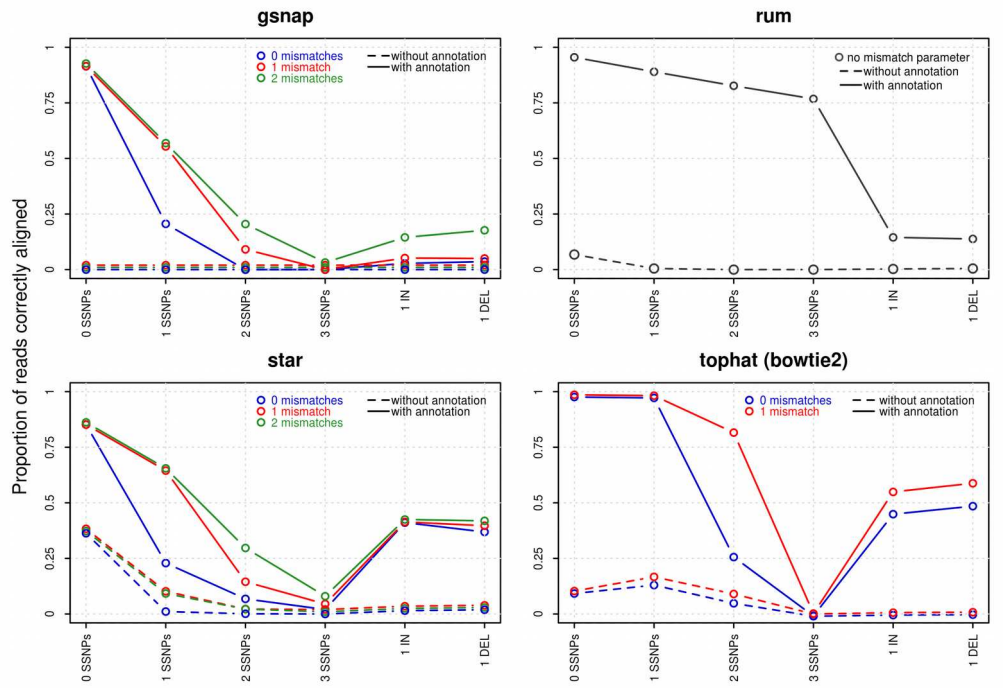


Figure 7. The accuracy of each short read alignment tool was evaluated for genome splice junction alignment using different IM/INDEL and mismatches metrics and with/without an annotation file (UCSC genes.gtf).

Genome splice junction alignment throughput. The throughput for each of the 4 tools for the alignment of 10 million simulated ribo-seq reads sampled across splice junctions, in the human genome, is provided in Figure 8. Interestingly, gsnap displayed the highest throughput. However, the throughput must be considered in the context of accuracy as discussed in the following section.

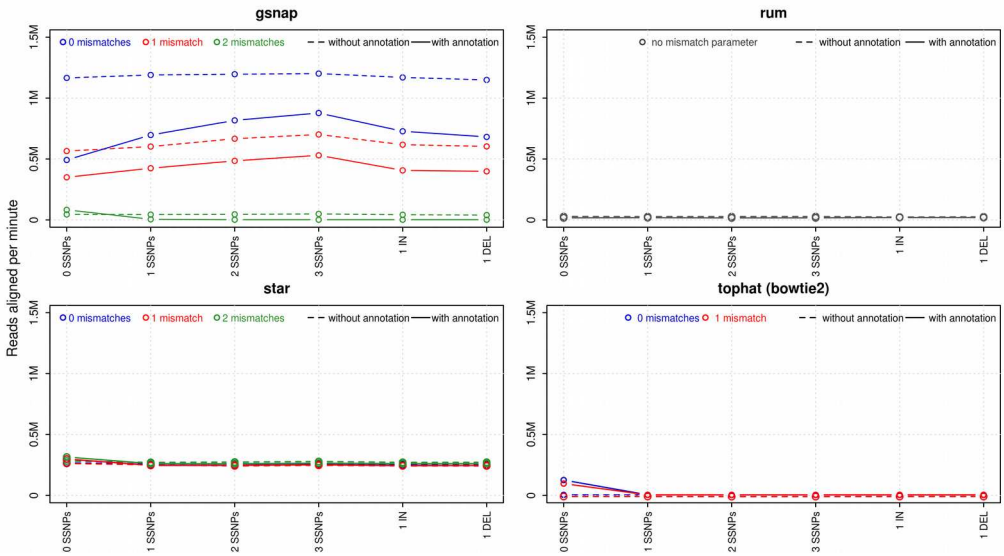


Figure 8. The throughput of each short read alignment tool was evaluated for genome splice junction alignment using different IM/INDEL and mismatch metrics and with/without an annotation file (UCSC genes.gtf).

Genome splice junction alignment accuracy and throughput. Although gsnap achieved the highest throughput for the simulated splice junction reads, its accuracy was lower than that of tophat (with bowtie2) and rum (Figure 9). Hence, the best short read alignment tool for splice junction mapping of ribo-seq data, would be tophat (with bowtie2), despite possessing a lower throughput than gsnap or star.

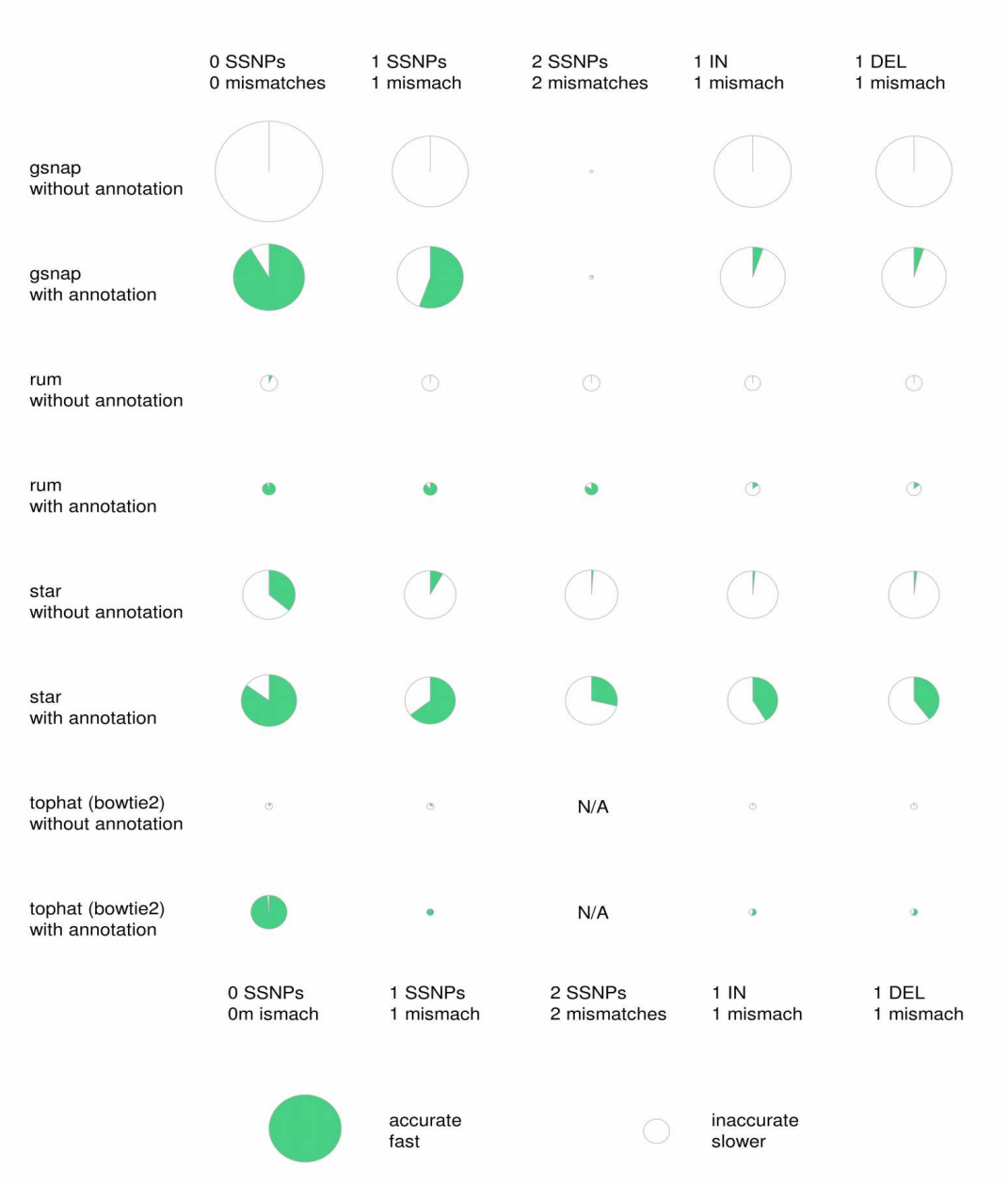


Figure 9. The accuracy and throughput of each short read alignment tool was evaluated for genome splice junction alignment using different IM/INDEL and mismatch metrics, and with/without an annotation file (UCSC genes.gtf). The size of the pie chart is proportional to the throughput achieved by the alignment tool, while the green slicing of the pie chart represents the alignment accuracy.

Genome splice junction alignment sensitivity and specificity. The number of TPs, FPs and FNs for each tool for splice junction benchmarking are provided in Figure 10. It is evident that, for all 4 tools, the number of TPs increased and the number of FPs and FNs decreased, when genome

annotations were supplied with the input alignment parameters. Overall, tophat (with bowtie2) had the highest number of TPs and the lowest number of FPs and FNs, with 0 IMs and 0 mismatches allowed and genome annotations used in the analysis.

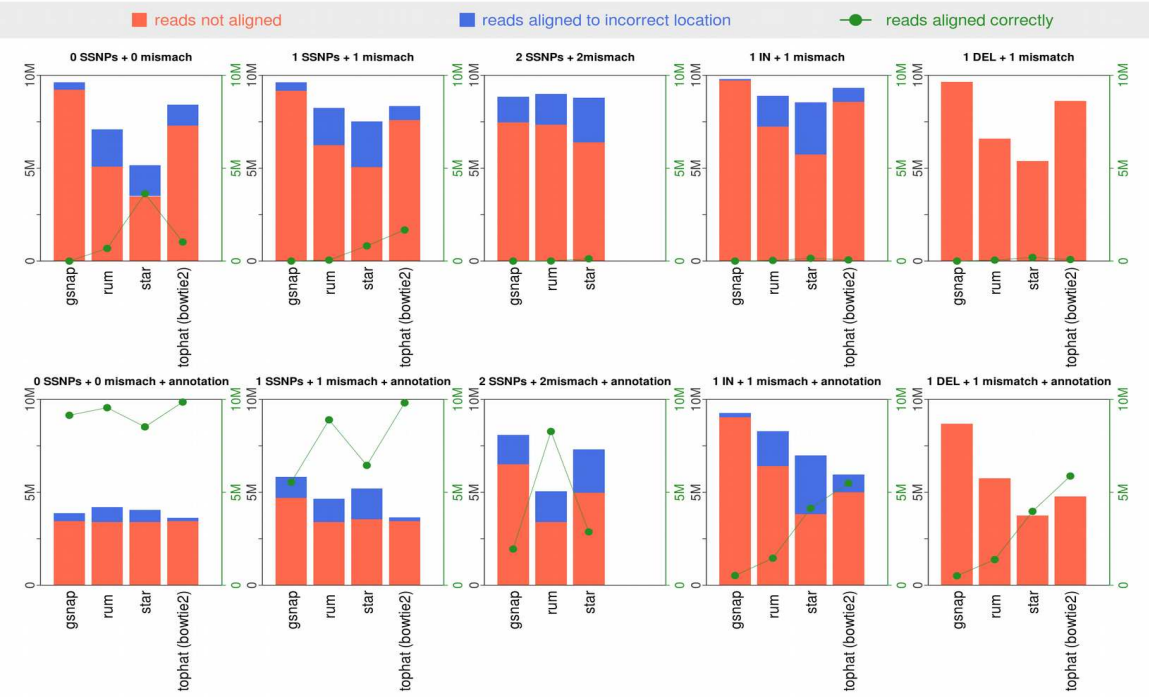


Figure 10. The true positives (TPs), false positives (FPs) and false negatives (FNs) of each short read alignment tool evaluated for splice junction alignment using different IM/INDEL, mismatch metrics and with/without an annotation file (UCSC genes.gtf). The green lines represent the number of reads out of 10 million, that were mapped to the original sampled location (TPs). The orange bars represent the number of reads not aligned (FNs) and the blue bars represent the number of reads aligned to another location, where the sequence distance was greater than the mismatch allowed metric.

4 Acknowledgment

Audrey Michel and Pavel Baranov are supported by Science Foundation Ireland [12/IA/1335]. Paul Walsh and Xiangwu Lu are funded researchers on the EU funded MetaPlat Project, NO. 690998.

5 References

Anon, UCSC Genome Browser, associated command-line software. Available at: http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/.

Anon, UCSC Genome Browser, File Server, hg19. Available at: http://hgdownload.cse.ucsc.edu/gbdb/hg19/ [Accessed March 8, 2009b].

Anon, UCSC Genome Browser, RefGene Annotation. Available at: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz [Accessed November 2, 2014c].

Dobin, A. et al., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England), 29(1), pp.15–21.

Grant, G.R. et al., 2011. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics (Oxford, England), 27(18), pp.2518–28.

Guydosh, N.R. & Green, R., 2014. Dom34 Rescues Riosomes in 3' Untranslated Regions. , 156(5), pp.950–962.

Hatem, A. et al., 2013. Benchmarking short sequence mapping tools. BMC bioinformatics, 14, pp.184.

Holtgrewe, M. et al., 2011. A novel and well-defined benchmarking method for second generation read mapping. BMC bioinformatics, 12(1), pp.210.

Homer, N., Merriman, B. & Nelson, S.F., 2009. BFAST: an alignment tool for large scale genome resequencing. PloS one, 4(11), pp.e7767.

Ingolia, N.T. et al., 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science (New York, N.Y.), 324(5924), pp.218–23.

Ingolia, N.T., 2014. Ribosome profiling: new views of translation, from single codons to genome scale. Nature reviews. Genetics, 15(3), pp.205–13.

Kent, W.J., 2002. BLAT--the BLAST-like alignment tool. Genome research, 12(4), pp.656–64.

Langmead, B., 2010. Aligning short sequencing reads with Bowtie. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], Chapter 11, Unit 11.7.

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), pp.357–9.

Lareau, L.F. et al., 2014. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. eLife, 3, pp.e01257.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 25(14), pp.1754–1760.

Li, H., Ruan, J. & Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research, 18(11), pp.1851–8.

Li, R. et al., 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics (Oxford, England), 25(15), pp.1966–7.

Marco-sola, S. et al., 2012. The GEM mapper? fast , accurate and versatile alignment by filtration. , 9(12).

Michel, A.M. & Baranov, P. V, 2013. Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. Wiley interdisciplinary reviews. RNA, 4(5), pp.473–90.

Ning, Z., Cox, A.J. & Mullikin, J.C., 2001. SSAHA: a fast search method for large DNA databases. Genome research, 11(10), pp.1725–9.

O'Connor, P.B.F. et al., 2013. rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. Bioinformatics (Oxford, England), 29(12), pp.1488–91.

Ponsting, H. & Ning, Z., 2010. SMALT - A New Mapper for DNA Sequencing Reads. , pp.1. Available at: http://cdn.f1000.com/posters/docs/327 [Accessed March 30, 2015].

Ruffalo, M., LaFramboise, T. & Koyutürk, M., 2011. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics (Oxford, England), 27(20), pp.2790–6.

Schbath, S. et al., 2012. Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. Journal of computational biology? a journal of computational molecular cell biology, 19(6), pp.796–813.

Steitz, J.A., 1969. Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. Nature, 224(5223), pp.957–64.

Trapnell, C., Pachter, L. & Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics (Oxford, England), 25(9), pp.1105–11.

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics, 10(1), pp.57–63.

Weiss, R.B. & Atkins, J.F., 2011. Molecular biology. Translation goes global. Science (New York, N.Y.), 334(6062), pp.1509–10.

Wu, T.D. & Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics (Oxford, England), 26(7), pp.873–81.

The Moorepark Grass Growth model a user-friendly model for farmers and researchers

E. Ruelle [1], D. Hennessy [1], L. Shalloo [1], P. Walsh [2], T. Manning [2], S. Wietrich [2] and L. Delaby [3]

[1] Animal & Grassland Research and Innovation Centre, Teagasc, Moorepark, Fermoy, Co. Cork, Ireland

[2] Department of Computer Science CIT – Cork Institute of Technology, Ireland

[3] INRA, AgroCampus Ouest, UMR 1348, Physiologie, Environnement and Génétique pour l'Animal et les Systèmes d'Elevage, F-35590 Saint-Gilles, France

elodie.ruelle@teagasc.ie

Keywords

Grass growth, model, graphical user interface

Abstract

In temperate regions, grazed grass is the most economical method of feeding ruminant livestock. Consequently, a model predicting grass growth as influenced by weather and grassland management would be an important tool for the farmer to improve his efficiency. The Moorepark Grass Growth model is a dynamic model for predicting grass growth, grass nitrogen content and soil nitrogen on a daily basis, based on weather conditions, soil type and grazing management. The model also takes into account the heterogeneity of a paddock which occurs due to deposition of animal excreta (urine and faeces). The numbers of inputs used by the model has been kept simple and minimal as possible and a user friendly graphical interface has been developed to allow the model to be used by all the actors in the agronomic sector. In this paper, the ability of the model to respond to different weather and nitrogen fertilisation strategies is examined. Overall, the model shows a good agreement with the previously known biological responses.

1 Introduction

In temperate climates, grass provides the cheapest and highest quality feed for dairy cows (Dillon et al., 2005). However, even though a temperate climate allows grass growth throughout the year, grass growth is highly seasonal and depends heavily on climatic conditions. This makes the volume of grass feed available variable and difficult to predict, and therefore difficult to factor into an efficient farm management strategy. Management of pasture (such as fertilisation, cutting and grazing severity) is also an important factor influencing grass growth. Nitrogen (N) fertiliser is highly effective at increasing grass growth and hence farm productivity, but it also contributes to increasing the risk of pollution of the groundwater through N leaching (Delaby et al., 1997). On a grazed paddock, urine and faeces depositions from grazing animals increases the N added through fertilization which increases the N availability for grass growth and the risk of N losses (leaching, volatilization) inconsistently across the paddock (Decau et al., 2004). Traditionally, soil models simulate urine patches uniformly across the whole paddock, losing the possible impact of the heterogeneity of the depositions. The Moorepark Grass Growth model (MGGm) described in this paper is a user friendly model capable of predicting grass growth and N utilisation on farm, that is able to model the heterogeneity of growth and N leached in a grazed paddock due to grazing animal depositions.

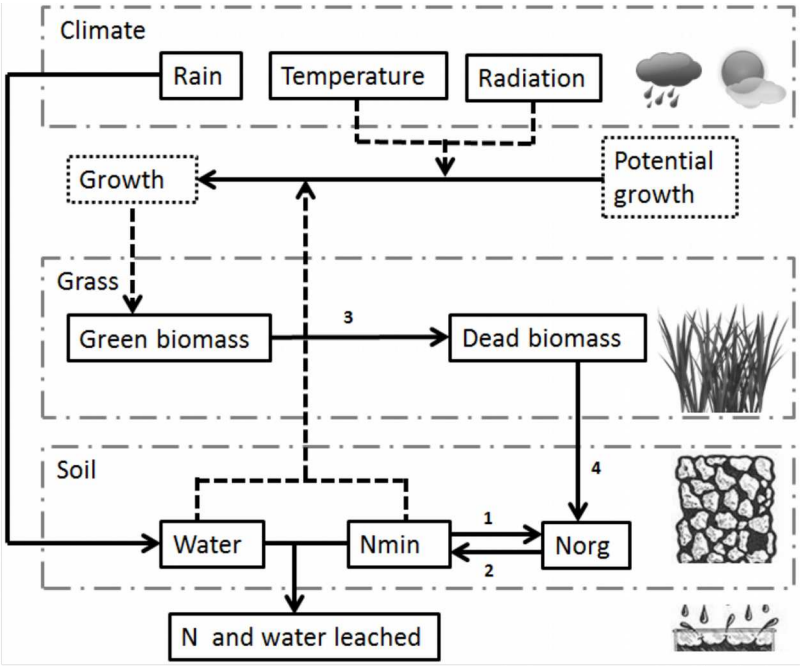
Decision support tools can provide farmers with information to increase grass growth and utilization. However to be practical for everyday farm use, models and model inputs have to be as

simple as possible and must be accessible through a user-friendly graphical interface, such as the one developed for the MGGm model and presented in this paper.

2 Material and Methods

2.1 Model presentation

The MGGm is a dynamic mechanistic model developed in C++ describing grass growth and N fluxes in a grass paddock at a 2 m2 scale. The model runs in a daily time step simulating soil N mineralisation, immobilisation, water fluxes, grass growth, N uptake and grass N content. In this model, each faeces and urine deposition is considered to affect only a 2 m2 area of the paddock. Therefore, the paddock is represented as a grid of 2 m2 having individual mineral N, organic N, grass biomass and grass N content. The localisation of each deposition on that grid is random with overlaps. The model is driven by a daily potential growth curve depending on the solar radiation and the total green biomass. To calculate the actual daily growth rate, this potential growth is multiplied by parameters depending on environmental conditions (temperature, water in the soil and solar radiation) and a parameter depending on the availability of the mineral N in the soil compared to the grass demand for N associated with the potential grass growth. The availability of the N in the soil depends on the mineral N in the soil, the proportion of the N usable by the plant (depending on the time of the year and the heading date), and the N demand to grow one kg of biomass. The N dilution curve represents the decrease of the N needed for one kg of dry matter (DM) growth with the increase of the accumulated biomass. A basic description of the biological model is presented in Figure 1.



Legend: (1) Immobilisation, (2) Mineralisation, (3) Senescence, (4) Abscission

Figure 1: Flow diagram of the main interaction of the model

The main inputs to the model are soil type (in terms of clay, sand and organic matter content), grassland management (including dates of grazing and cutting events, number of animals, post grazing/cutting height), fertiliser management (date and quantity of N applied), and daily weather

(in terms of minimum, maximum and average temperature, rain and radiation). The main outputs from the model are the daily grass growth (kg DM/hectare (ha)/day) and biomass (kg DM/ha), N leaching (kg N/ha), the N organic and mineral content in the soil (kg N/ha) and the N content of the grass (g N/kg DM). Those outputs are presented at both the paddock and the 2 m2 level allowing users to understand the impact of urine and faeces deposition.

2.2 Development of the graphical interface

To make the model accessible to both researchers and famers, a graphical user interface (GUI) for the model is developed to simplify its use and provide active feedback for non-technical users, but which also provides efficient data entry and experiment design functionality for power-users. A screenshot of the GUI developed is shown in Figure 2.

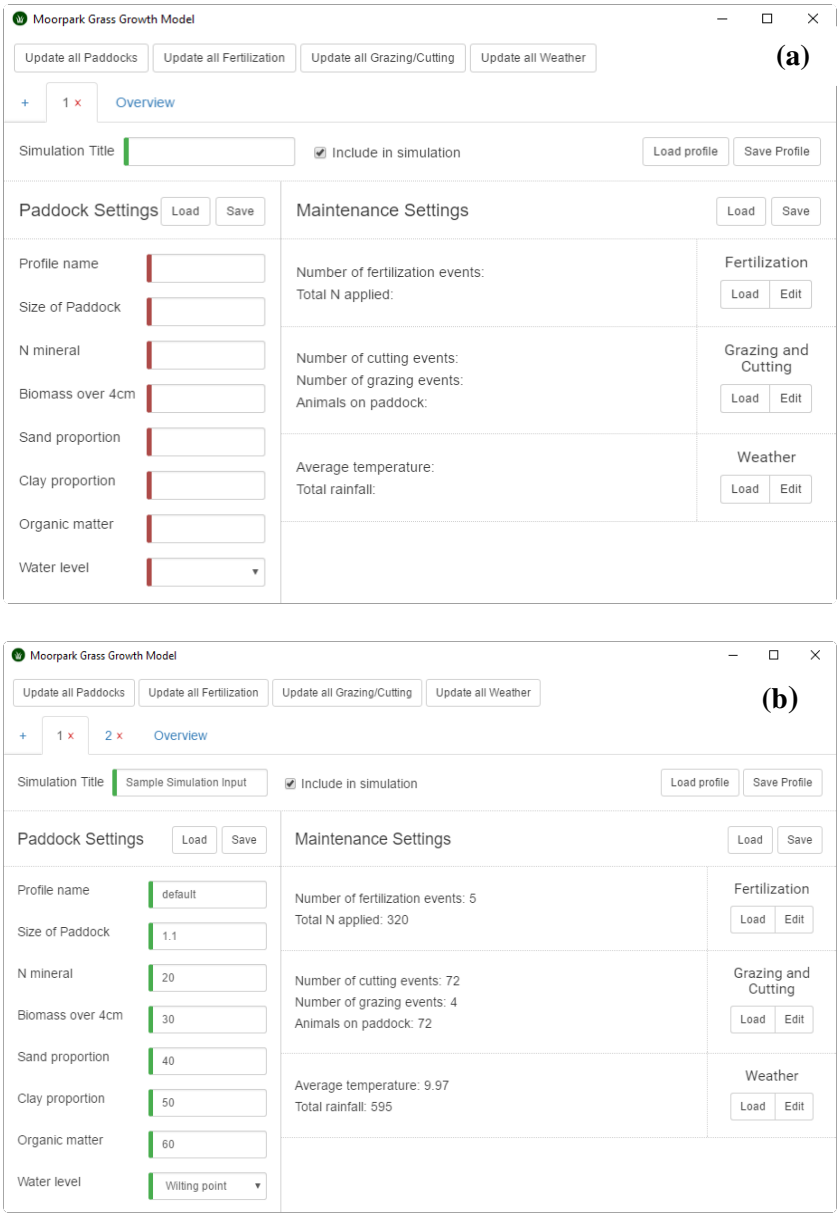


Figure 2: (a) The GUI developed for the MGGm system showing a simulation template which must be completed, and (b) the MGGm with a valid simulation configuration loaded

To organize multiple simulations to run at the same time, a tab based interface is implemented where each simulation (set of inputs for paddock, grazing management, and weather) is presented on a separate tab. Tabs can be added or deleted to change the number of simulations to be run. There is also functionality to update specific elements across all tabs. The simulations in a single run could be, for example, the same paddock with different grazing management, or simulations could be across different paddocks with the same grazing management, or investigating the impact of different fertiliser and grazing regimes, or even comparing the grass growth using the same management given the weather experienced in different years. As information from different simulations is hidden from the user as much as possible, an additional "overview" tab is included which presents a table based summary of the information on all tabs.

The GUI for each simulation is divided into different sections; paddock; fertilization, grazing/cutting and weather. As the paddock section requires a set number of inputs, text boxes are provided for each element that prevent nonsensical input and alert the user to missing necessary values or values outside of the expected ranges. For the management regimes and weather sections, large amounts of data can be required. For these sections, the user has the option to manually edit the values on the interface, or they are provided with clearly structured template files that can be edited using standard spreadsheet software, as this is the most typical and efficient approach to dealing with this sort of data. Validation for this data is provided in the form of summary statistics for each file that can highlight erroneous or invalid data to the user.

To further increase the efficiency of the system, sets of simulation inputs can be saved and loaded, so that they can be reused or modified at a later stage, as there is a lot of overlap in runs in the expected use cases. Here, the entire simulations, the paddocks, the grazing/cutting events, the fertilization events, the weather, or a combination of grazing/cutting events, fertilization events and weather can be saved as a "profile" for future runs.

For the power users, the system outputs spreadsheets containing a detailed analysis of the paddock conditions across a range of useful metrics on a day by day basis. The typical farmer however will only be interested in the levels of grass growth and biomass, and as such, the system is able to generate easily interpretable graphs of these values, as shown in Figure 3.

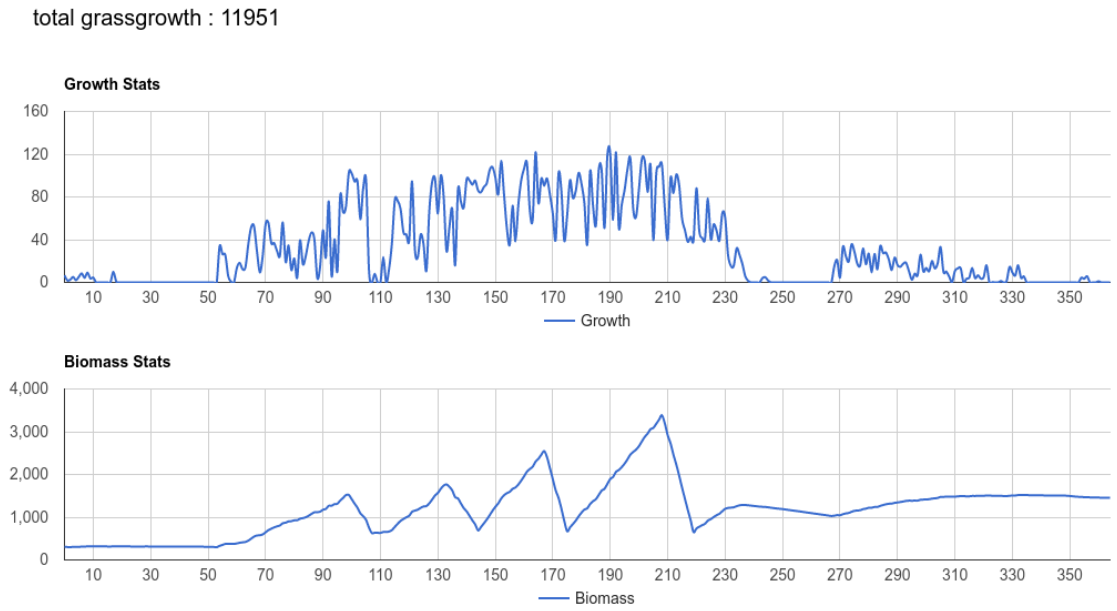


Figure 3: Example graphical output generated by the system

2.3 Evaluation of the model

Different simulations have been run to evaluate the impact of different mineral N fertiliser strategies on grazing conditions across different weather years. These simulations have been tested using the weather conditions from Ireland (Co. Cork) and France (Normandy), described in Table 1. Mineral N applications were tested at 0 (no fertilisation 0F) or 200 kg of N per ha per year in Ireland (with fertilisation F). Due to the high temperature and lack of rain in summer in Normandy, the annual mineral N fertiliser application was reduced by 25% through the removal of autumn fertiliser application. Paddock size was 1 ha and a total of seven (France) or eight (Ireland) grazing events were simulated. Grazing was simulated on the 78nd, 113th, 134th, 159th, 189th, 265th and 305th day each year for both countries, and additionally on the 225th day in Ireland, in a 365 day year. The initial N mineral status of the soil was 80 kg N/ha and the N organic level was at 14,400 kg N/ha (6% of soil organic matter content). In the simulations, the number of animals for each grazing event was calculated based on the pre-grazing sward height (calculated by the model) and a daily animal DM intake of 16 kg DM.

Location Year	Annual average temperature (°C)	Monthly rainfall (mm)												Annual total rainfall (mm)
		J	F	M	A	M	J	J	A	S	O	N	D	
France Normandy	10.2	61	32	136	68	64	12	35	72	42	80	102	41	743
Ireland Co.Cork	8.7	107	39	88	59	38	53	143	23	102	83	98	37	869

Table 1: Description of the two annual weather datasets applied in the simulations

The simulations were completed across two years with the same weather and management conditions for each location allowing the quantification of the impact of the first grazing year on the leaching that occurred during the winter and spring of the second year. The N leaching results are presented between the 1st of November of year one and the 30th of April of year two (called the winter period). All the other results concern only the first grazing year of the simulation. To highlight the heterogeneity within a paddock for the 2 m2 simulation, results are presented in the form of minimum (Min), maximum (Max) and standard deviation (SD). The Min and Max represent the 2 m2 with the lowest or highest value of the variable considered, the SD represents the standard deviation of those values within a paddock. Due to the variables considered (net growth, N mineral and N leached), the Min for each variable corresponds to an area of 2 m2 which received no urine or faeces deposition.

3 Result and discussion

3.1 Impact of the weather and nitrogen fertilisation

Table 2 describes the variability of grass growth across the year, the mineral N content in the soil on the first of November and N leached during the winter period within a paddock. Figure 4 plots the net grass growth (in terms of kg DM) calculated by the model for the first year of the simulation in both Ireland and France both with and without the use of fertilizer. The average grass grown by the 0F simulation was 7,336 kg DM/ha based on the French weather and 9,146 kg DM/ha with the Irish weather. In comparison the F simulation grew to 9,350 kg DM/ha for France and 12,777 kg DM/ha

for Ireland (Table 2). This shows that the model is capable of reacting to different weather conditions and is highly sensitive to different levels of mineral N fertilisation (Figure 4). The lack in rainfall in France in 2008 led to poor grass growth in summer and autumn. The response to N fertiliser application rate is comparable with published figures (Whitehead, 1995).

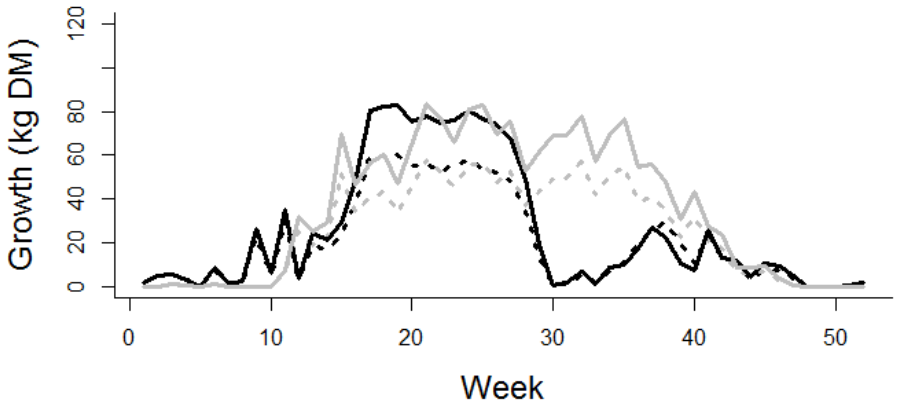


Figure 4: Net growth for the French 2008 weather (black) and the Irish 2009 weather (grey) for the 0F (dotted line) and F (continuous line) simulations.

	Annual grass grown (kg DM/ha)		N uptake (kg N/year)		Number of grazing days (/ha/year)		N Urine (kg per year per ha affected)		Soil mineral N (kg N/ha) content on 1st Nov		N leached (kg N/ha)	
Fert	0F	F	0F	F	0F	F	0F	F	0F	F	0F	F
France Normandy	7,336	9,350	139	237	376	501	98[1]	153[2]	97	151	37	52
Ireland Co. Cork	9,146	12,777	169	327	524	758	88[3]	171[4]	103	159	61	91

[1]On 52% of the paddock [2]On 63% of the paddock
[3]On 65% of the paddock [4]On 78% of the paddock

Table 2: Impact of the weather and the N fertilisation on total annual grass grown (kg DM/ha), total annual N uptake (kg N/ha), the number of grazing days per ha, the N from the urine of the animal, the soil mineral N (kg N/ha) and the total annual N leached (kg N/ha).

The application of N fertiliser resulted in an increase in grass growth under both weather conditions leading to an increase in the number of grazing days. Depending on the number of grazing days, the proportion of the paddock affected by urine deposition was 52%, 63%, 65% and 78% for the 0F French simulations, the F French simulations, the 0F Irish simulations and the F Irish simulations, respectively. The application of N fertiliser also led to an increase in the quantity of N leached in the soil. The N leached in Ireland was higher than in France probably due to a higher N mineral content of the soil, a higher number of grazing days, and higher rainfall.

3.2 Variability due to the localised simulation

The impacts of the localised animal excreta depositions are presented in Table 3. The accumulation of urine patches result in an increase in grass growth of between 3,536 and 5,052 kg DM/ha in the simulations. Similarly, the N leaching differs between the different 2 m2 areas with a maximum difference in the F Irish simulation (increase of 134 kg N leached/ha). The lowest maximum difference in N leached is in the 0F French simulation with a difference of 33 kg N leached/ha.

		Annual grass grown (kg DM/ha)		Soil mineral N (kg N/ha) content on 1st Nov		N leached (kg N/ha)	
N (France / Ireland)		0F	F	0F	F	0F	F
France Normandy	Min	6575	8544	86	119	34	45
	Max	10909	12080	201	454	67	116
	SD	829	732	15	39	4	9
Ireland Co.Cork	Min	8226	11564	91	125	54	72
	Max	13278	16573	194	411	102	206
	SD	901	902	13	35	7	18

Table 3: Internal heterogeneity of the paddock grazing condition. The maximum (Max), minimum (Min) and standard deviation (SD) of the annual grass grown (kg DM/ha), soil mineral N content (kg N/ha) and the total N leached (kg N/ha) at the 2 m2 level within a paddock.

3.3 Graphical interface

The graphical interface developed for this model facilitates greater penetration at both farm and research levels. There are two main applications envisioned for the new user friendly tool depending on the end user. The farmer will generally use it to predict changes in grass growth in the following week to 10 days, and compare different management strategies, while the researchers will use it to investigate the effect of different management regimes on various environmental and economic outputs. Therefore, the system must be straight forward enough to be used by non-technical users, while being efficient enough for power users and researchers to use extensively with minimal frustration.

4 Conclusion and future work

The MGGm is a dynamic model permitting the simulation of different weather conditions and grazing management on the grass growth profile. Due to its graphical interface and the small number of easily accessible inputs, this model can be used by any actor in the dairy system sector. Despite its level of precision (2 m2 level leading to 5,000 different measurements of grass growth, biomass, etc. for a 1 ha paddock per day) the model is not slow (less than 2s for a 365 day simulation of a 1 ha paddock).

The model is in the process of being evaluated against actual data and shows good accuracy. The model has been integrated into a whole farm model (Ruelle et al., 2015), permitting a precise simulation of grass growth on each paddock of the farm. In the future, the model will be implemented into PastureBase Ireland (Griffith et al., 2013), and when linked with localised

weather forecast it will provide the farmer with a prediction of grass availability on his farm in the next week to 10 days.

5 Acknowledgment

The authors acknowledge the funding from the Research Stimulus Fund 2011 administered by the department of Agriculture, Food and the Marine (Project 11/S/132).

6 References

M. Decau, J. Simon and A. Jacquet (2004), "Nitrate leaching under grassland as affected by mineral nitrogen fertilization and cattle urine," in: Journal of Environmental Quality, vol. 33, no. 2, pp.637-644.

L. Delaby, M. Decau, J. Peyraud, and P. Accarie (1997), "AzoPât: une description quantifiée des flux annuels d'azote en prairie pâturée par les vaches laitières. I. Les flux associés à l'animal", in: Fourrages, vol. 151, pp. 297-311.

P. Dillon , J. R. Roche, L. Shalloo, and B. Horan (2005), "Optimizing financial returns from grazing in temperate pastures," in: Proceedings of the XX International Grassland Congress, Cork Satellite, 04-Jul-2005, pp. 13.

V. Griffith, A. Geoghegan, M. O'Donovan, and L. Shalloo (2013), "PastureBaseIreland-National grassland database," in: Proceedings of Moorepark'13 Irish dairying, harvesting the potential, Fermoy, Ireland, pp: 60-6.

E. Ruelle, L. Shalloo, M. Wallace, and L. Delaby (2015), "Development and evaluation of the pasture-based herd dynamic milk (PBHDM) model for dairy systems," in: European Journal of Agronomy vol: 71, pp: 106-114.

D. C. Whitehead (1995), Grassland nitrogen, CAB international.

A mHealth platform for supporting self-management of chronic conditions

Huiru Zheng[1], Timothy Patterson[1], Chris D. Nugent[1], Paul J. McCullagh[1] , Mark P. Donnelly[1], Ian Cleland[1] , Suzanne McDonough[2] , Adele Boyd[1], Norman D. Black[1]
[1] School of Computing and Mathematics, [2] School of Health Sciences
University of Ulster, Shore Road, Jordanstown, Newtownabbey, Co. Antrim, BT37 0QB
e-mail: {h.zheng, t.patterson, cd.nugent, pj.mccullagh, mp.donnelly, i.cleland, a.boyd, s.mcdonough, nd.black}@ulster.ac.uk

Keywords: self-management, mHealth, chronic conditions

Background

Self-management is an crucial approach for people with chronic conditions in daily management of their conditions. Mobile health (mHealth) has been used to an effective tool for self-management of various chronic conditions, for example, stroke, chronic pain, diabetics and early dementia. It can provide basic information about the related disease, set up individual's self-management goal, assist in adopting a self-management care plan, lifestyles and intervention, and monitoring the progress in reaching the goal. However, most of mHealth applications are specific designed and lack of scalability. In this paper we present our current work in the development of a generic self-management mHealth platform, KeepWell, which can be readily extended for specific chronic conditions. The evaluation of KeepWell on self-management of the chronic obstructive pulmonary disease (COPD) will be presented and how our proposed design may be adapted for a specific condition will be demonstrated. The project is funded by Invest Northern Ireland (RD0513844).

Methods

The design of the generic mHealth platform KeepWell is built upon previous research on home based self management and is in collaboration with clinical experts. Functional requirements for the KeepWell platform align with the National Institute for Clinical Excellence (NICE) guidelines on COPD management (NICE, 2010) and Stroke Management (NICE, 2013) and the conceptualization of self-management for early stage Dementia by Martin et al. (2012) and can be summarised into five main functions of the self-management, i.e. activity planning, self-reporting, educational information, facilitation of social engagement and device measurement. The platform consists of four main components (Patterson et al, 2014): (1) an Android app that facilitates consumption of educational material, goal setting and display of health metrics; (2) health-tracking devices that enable the capture and quantification of measurements where state-of-the-art off-the-shelf devices such as smart wristband, wireless blood pressure monitor and wireless scales are used; (3) servers that provide storage and perform computation; and (4) a web-based portal. The goal setting is personalised according to each individual's condition and health status. This can be set with the support from clinicians. The tracking devices can easily be extended or added. The web-based portal enables the management of educational resources and self-report including the symptoms and workouts. These four components realise the five main functions of the self-management, i.e. activity planning, self-reporting, educational information, facilitation of social engagement and device measurement.

Results

The system has been developed and implemented. Currently a suite of Witherings devices are used for healthcare measurements. A single testing phase of a beta version of the platform has been conducted by members of the project team (N = 9) for approximately one week. After that, the usability of the KeepWell mobile application for COPD was tested by eight clinicians on a one off occasion.

Clinicians were contacted by one of the research team to arrange a time and date for an individual meeting. The meetings took place across Northern Ireland at the clinician's place of work or at the Ulster University. Before the participants began, the procedure was explained fully and they had a chance to ask questions. Clinical opinion was obtained by asking eight clinicians (six working in the area of COPD management across Northern Ireland; and two working in a research environment) to complete a series of tasks on KeepWell (see Table 1 below). Throughout the tasks, clinicians were permitted to consult with a printed manual, which explained the operation of the app, as needed.

Task	Purpose
Clinician to complete the following steps • Put activity tracker together and position on wrist. • Launch app on mobile phone device. • Set a step goal. • Self-report a workout or Borg. • View progress within app including daily/weekly. • Create a reminder using date and time. • Change/edit reminder date and time.	To demonstrate how to wear and use the wearable device. To demonstrate overall function of app and highlight specific features. To obtain clinician opinion on usefulness of features.
Clinician to view educational materials	Check format of the included materials.
Clinician to review suitability of self-report workouts included in KeepWell.	Check suitability of workouts included.
Clinician to complete questionnaire.	Quantitative and qualitative feedback

Table 1. *Tasks completed during the clinical evaluation. PR=pulmonary rehabilitation.*

Ethical approval was not required as the study was conducted under the auspices of Personal and Public Involvement that is the involvement of expert clinician as specialist advisers in the design of the KeepWell app.

The clinicians were asked to verbalise what they thinking about, looking at, doing and feeling throughout the process. They provided valuable knowledge and expertise based on their experience of COPD management in the development of the KeepWell platform. No information was

collected that could identify the clinicians involved, as the sole purpose was to evaluate usability of the KeepWell app.

After the 'think aloud' study had finished, users were asked if they have any suggestions for how the app could be improved or any additional comments they wished to make and were asked to complete a usability questionnaire. Any qualitative comments or any problems during the completion of each task were noted.

Overall the participants reported high levels of usability of the app and the suggestions have informed the improvement of the design (McDonough et al, 2016).

Summary

The KeepWell self-management mHealth platform generically implemented to accommodate three chronic conditions, and it is extensible to goal-setting, measurement, educational information, self-reporting and feedback components are generically implemented to accommodate three chronic conditions. The further work will be carried out to improve the usability of the KeepWell for COPD self-management and testing of KeepWell for stroke and dementia self-management.

References

National Institute for Clinical Excellence (NICE) (2010). Chronic Obstructive Pulmonary Disease Management of Chronic Obstructive Pulmonary Disease in Adults in Primary and Secondary Care. Technical Report April 2007, NICE, 2010.

National Institute for Clinical Excellence (NICE) (2013). Stroke Rehabilitation Long-Term Rehabilitation after Stroke. Technical Report April 2007, NICE, 2013.

S. M. McDonough, A. Boyd, T. Patterson, P. McCullagh, I. Cleland, C. Nugent, M. Donnelly, H. Zheng and N Black(2016), Development of mobile phone app in collaboration with Chronic Obstructive Pulmonary Disease clinical experts, ICDVRAT2016, to be presented.

T. Patterson, I. Cleland, C.D. Nugent, N.D. Black, P. McCullagh, H. Zheng, M.P. Donnelly, and S. McDonough (2014), Towards a Generic Platform for the Self-Management of Chronic Conditions. In Bioinformatics and Biomedicine (BIBM), 2014, IEEE International Conference on, pages 44–47, Belfast, Northern Ireland, 2014. IEEE.

Development of a Bioinformatics Pipeline to Identify Antibiotic Resistance Biomarkers of Bacterial Pathogens

D. Mullins [1,2], N. Konstantinidiou [1], R.D. Sleator [3], B. Kelly [1], T. Forster [4], W. A. Dantoft [4], P. Ghazal [4], P. Walsh [1*]

[1] NSilico Life Science Ltd., Cork

[2] School of Microbiology, University College Cork, Cork, Ireland

[3] Department of Biological Sciences, Cork Institute of Technology, Bishopstown, Cork, Ireland

[4] Division of Infection and Pathway Medicine, Edinburgh Medical School, Edinburgh, United Kingdom

[*] paul.walsh@nsilico.com

Keywords

Bioinformatics pipeline; pathogen biomarker; neonatal sepsis.

Abstract

Neonatal sepsis is a severe infection with a high mortality rate of newborns worldwide. Currently, standard treatment of bacterial sepsis includes antibiotic prescription. However, the rapid evolution of antibiotic resistance in sepsis pathogens foreshadows the urgent need for new therapies. Therefore, this research focused on the development of a software pipeline allowing in silico analysis of bacterial pathogens, to identify antibiotic resistance biomarkers. Biomarkers are useful for rapid establishment of antibiotic resistance genes and mechanisms, allowing prescription of an appropriate antibiotic that would greatly reduce neonatal sepsis mortality rates.

To develop an antibiotic resistance biomarker pipeline, sequenced genomes of two diverse bacterial agents of sepsis, Gram-positive L. monocytogenes EGD-e and Gram-negative S. enterica Typhi 14028, were studied in silico. The input bacterial genome data were sequenced with an Illumina MiSeq desktop sequencer and stored in FASTQ file format. To analyse the bacterial genomes, several bioinformatics tools were reviewed and command-line tools were preferred due to their high potential for automated processes.

The overall biomarker identification pipeline developed here, was broken into two consecutive but separate parts comprising 'Genome assembly and annotation' and 'Comparative genomics and biomarker identification' sub-pipelines. The function of the first sub-pipeline included the assembly and annotation of a bacterial genome from raw sequence data. The second sub-pipeline provided a homology-based comparative genomics analysis, which required several annotated genomes as input. The comparative analyses incorporated several approaches such as phylogenetic tree construction, resistome profile investigation and pan-genome study. Biomarkers were discovered through the above techniques, after comparison of the pathogens with each other as well as with non-pathogenic strains. The intelligence on the genetic secrets of a pathogen could be utilised by a clinician or a researcher to improve the existing treatments or to develop more effective therapeutic strategies. The biomarker identification pipeline possesses the potential for the study of any bacterial pathogen. However, further development is required for full automation of this pipeline.

1 Introduction

Sepsis is a severe and complex infection, with particularly high mortality rates in neonates, often caused by bacteria (Simonsen et al., 2014; Liu et al., 2015). In immature immune system of newborns, bacteria can provoke uncontrolled immune system activation (sepsis) that can rapidly lead to organ failure and death (Smith et al., 2014). Currently, antibiotics and intensive hospital care include the main treatment strategies (Simonsen et al., 2014). However, the rapid in silico identification of pathogen biomarkers can aid clinicians in selecting the correct course of treatment, and thus reduce the sepsis mortality rates (Reinhart et al., 2012). This research aimed to achieve this goal by the development of a software pipeline to identify pathogen biomarkers associated with antimicrobial resistance and virulence factors.

To develop the pathogen biomarker identification pipeline, two different sepsis pathogens, the raw sequences of Gram-positive bacterium Listeria monocytogenes (Glaser et al., 2001; Dickinson et al., 2015) and Gram-negative Salmonella enterica serovar Typhi (McClelland et al., 2001; Stoesser et al., 2013) were utilised in this research. The term virulence factor, as used here, refers to the genes of a pathogen that contribute to its disease-causing ability (Niu et al., 2013).

Virulence factors in both pathogens, Listeria and Salmonella, can induce or repress the innate inflammatory response (Rathinam et al., 2012). An example of this is the downregulation of flagellin, which enables Salmonella species to avoid their detection by the immune system (Rathinam et al., 2012). This action of the pathogens, combined with the immature immune system in neonates, result in a deadly combination (Smith et al., 2014).

Raw sequencing data of a bacterial genome is the prerequisite to the pipeline developed in this research. The raw genome sequencing data must first be computationally assembled and annotated for further downstream analysis. These analyses require significant bioinformatics skills. To avail reliable analyses, genome assembly and annotation are automated within NSilico's Simplicity™ platform (Walsh et al., 2013). The appeal of this pipeline lies in its ability to simplify and accelerate bioinformatics analyses for the researchers and clinicians that need it (Walsh et al., 2013).

Assembling a genome from its raw sequence reads involves joining overlapping regions to form long contiguous sequences known as contigs. Ekblom and Wolf (2014) suggested that de novo assembly, which does not make use of a reference genome, is preferred initially. However, to refine and extend these contigs into scaffolds, reference-assisted genome assembly may be employed, if appropriate reference genomes are available for the organism under investigation (Ekblom and Wolf, 2014). Two of the most popular de novo genome assemblers are Velvet and SPAdes (Zerbino, 2010; Bankevich et al., 2012). They are open-source tools, designed for assembling prokaryotic genomic sequencing data. ABACAS (Assefa et al., 2009) and CONTIGuator (Galardini et al., 2011) are reference-assisted genome assembly tools. They attempt to construct a single scaffold out of a set of contigs, by aligning them to a single reference genome. An alternative reference-assisted assembler is Ragout (Kolmogorov et al., 2014). To resolve the contigs into longer scaffolds, Ragout takes the novel approach of using multiple reference genomes and the phylogenetic relationship between them.

After genome assembly, the next step usually comprises a genome annotation process, during which putative genes are associated with biological functions. Preceding functional annotation, genes are ab initio predicted from the genome sequence. Prokka (Seemann, 2014) and RASTtk (Brettin et al., 2015) are two command-line tools that predict and annotate genes. Both tools aim to annotate the whole bacterial genome. Genes of special interest can also be annotated using different tools and databases. PHASTER is a database resource for the annotation of prophage sequences in bacteria (Arndt et al., 2016). Fortier and Sekulovic (2013) defined a prophage as the DNA of a bacteriophage that has integrated into the chromosome of a bacterium. Prophage genetic material encoding potent toxins can influence the course of evolution in pathogenic bacteria (Fortier and Sekulovic, 2013). Antibiotic resistance genes discovered in a pathogen represent a subset of the resistome. Wright (2007) defined the resistome as the entire set of antibiotic resistance genes present in both pathogenic and non-pathogenic bacteria. The Comprehensive Antibiotic Resistance Database (CARD) is the leading resource for antibiotic resistance genes (McArthur et al., 2013). Annotation of the antibiotic resistance genes can contribute in selecting the correct antibiotics required for treatment. Virulence factors are another group of genes required to study pathogenicity in sepsis bacteria. The Virulence Factor Database (VFDB) provides a comprehensive database which can be used to annotate genes as potential virulence factors (Chen et al., 2016).

The assembled and annotated genomes can then be compared to that of related well-sequenced organisms or non-pathogenic strains. Comparative genomics is a very useful field of study for inferring information about a genome. Comparing closely related non-pathogenic bacteria to a pathogenic bacterium under study, reveals important insights into the pathogenicity of the virulent bacterium. The phylogeny of an organism combined with lifestyle information is an important starting point, which can give extensive interpretability (Holden et al., 2013). Numerous genetic entities such as highly conserved 16S rRNA genes (Case et al., 2007), or the core-genome from a pan-genome analysis (Holden et al., 2013), are frequently used for the construction of a phylogenetic tree. A popular, command-line phylogenetic tool is RAxML (Stamatakis, 2014), which implements the maximum likelihood method for phylogenetic analyses. An additional comparative genomic step is to analyse the pan-genome, the "entire genomic repertoire of a given phylogenetic clade" as defined by Vernikos et al. (2015). The pan-genome may be sub-divided into the core-genome, shared by all the members of the clade, and the accessory genome, containing genes shared by a subset of organisms under investigation (Vernikos et al., 2015). The Roary tool (Page et al., 2015) can create a pan-genome using the output of the Prokka annotator (Seemann, 2014). Annotation of all genomes with Prokka provides a uniformity of annotation, hence preventing bias in the pan-genome analysis.

2 Materials and Methods

2.1 Bacterial Strains and Summary

This research aimed at automation of an in silico pipeline for biomarker identification initiating with raw sequence data of a bacterial pathogen. The pathogens under investigation were L. monocytogenes EGD-e and S. enterica Typhi 14028. L. monocytogenes EGD-e was originally isolated and studied by Glaser et al. (2001), before being sequenced by Illumina (2014) using their Illumina MiSeq® desktop sequencer (http://www.ncbi.nlm.nih.gov/sra/ERX705166%5Baccn). Alexander et al. (2016) isolated and sequenced S. enterica Typhi 14028, using the same sequencer technology (http://www.ncbi.nlm.nih.gov/sra/SRX1012435[accn]). The sequencing reads for both pathogens were stored in two FASTQ format files, with the first file corresponding to forward and the second to reverse reads. The collection of tools used during this research are summarised in Table 1.

2.2 Genome Assembly

The raw sequencing reads of L. monocytogenes and S. enterica were used to compare the two de novo assemblers, Velvet (Zerbino, 2010) and SPAdes (Bankevich et al., 2012). Velvet was utilised through its wrapper script, VelvetOptimiser, which optimises de Bruijn graph construction (Zerbino, 2010). The quality of genome assemblies and the set of contigs output by Velvet and SPAdes were evaluated using the QUAST tool (Gurevich et al., 2013). To calculate comparative statistics, a reference genome sequence and annotation files were provided to QUAST. The best reference genomes were selected by BLAST searches of the RefSeq reference genomes database (Tatusova et al., 2015). The reference genomes for each pathogen were L. monocytogenes EGD-e or S. enterica Typhi 14028S.

The de novo assembled contigs of L. monocytogenes and S. enterica, using SPAdes, were refined by methods of reference-assisted genome assembly. ABACAS (Assefa et al., 2009) and CONTIGuator (Galardini et al., 2011) were employed to assemble the genomes using the single best reference genome as a guide. The Ragout tool (Kolmogorov et al., 2014) is different from the

above assemblers since it requires multiple reference genomes for the improved assembly. The L. monocytogenes strains for Ragout assembly were EGD-e, FSL_R2-561, FW040025, Lm_3136, SLCC2372 and SLCC2479. The S. enterica Typhi strains included 14028S, LT2, 08-1736, UK-1 and VNP20009. All reference-assisted assemblies were compared using the QUAST tool as before.

Method	Tool	Reference
De novo genome assembly	Velvet v1.2.10 through VelvetOptimiser v2.2.5	(Zerbino, 2010)
	SPAdes v3.8.0	(Bankevich et al., 2012)
Quality assessment for genome assemblies	QUAST v4.1	(Gurevich et al., 2013)
Reference-assisted genome assembly	ABACAS v1.3.1-3	(Assefa et al., 2009)
	CONTIGuator v2.7	(Galardini et al., 2011)
	Ragout v1.2	(Kolmogorov et al., 2014)
Automatic genome annotation	RASTtk v1.3.0	(Brettin et al., 2015)
	Prokka v1.11	(Seemann, 2014)
Database searching	BLAST+ v2.2.31	(Camacho et al., 2009)
	HMMER v3.1b2	(Eddy, 2011)
Prophage gene annotation	PHASTER January 2016	(Arndt et al., 2016)
Antibiotic resistance gene annotation	CARD v1.0.8 Resistance Gene Identifier Tool v3.0.9	(McArthur et al., 2013)
Virulence factor annotation	VFDB core dataset 12/8/16	(Chen et al., 2016)
Phylogeny – Find 16S rRNA genes	RNAmmer v1.2	(Lagesen et al., 2007)
Phylogeny – Consensus sequence	Consambig tool within EMBOSS software suite v6.6.0.0	(Rice et al., 2000)
Phylogeny – Multiple alignment	MUSCLE v3.8.31	(Edgar, 2004)
Phylogeny – Inferring the phylogenetic tree	RAxML v8.2.9	(Stamatakis, 2014)
Phylogeny – Viewing the phylogenetic tree	Dendroscope v3.5.7	(Huson and Scornavacca, 2012)
Pan genome analysis	Roary v3.6.0	(Page et al., 2015)

Table 1: Summary of the methods and tools used to carry out in silico analyses of S. enterica and L. monocytogenes. The tools are in a sequential, logical order for a pipeline. The tool names are hyperlinks to their web pages.

2.3 Genome Annotation

The de novo assembled genomes of L. monocytogenes and S. enterica, derived from SPAdes and refined with Ragout, were now ready for annotation. RASTtk (Brettin et al., 2015) was initially used to ab initio predict genes. The predicted genes were then annotated for antibiotic resistance, prophage and virulence factor genes. Prokka (Seemann, 2014) was later employed due to its compatibility with the Roary pan-genome tool (Page et al., 2015). Prophage genes were annotated by BLASTP searches (Camacho et al., 2009) of the PHASTER database (Arndt et al., 2016). Antibiotic resistance genes were annotated from the CARD database using the provided Resistance Gene Identifier Tool (McArthur et al., 2013). Virulence factors were annotated via mining VFDB (Chen et al., 2016) with BLASTP.

2.4 Comparative Genomics

To draw biologically meaningful conclusions on the nature of the investigated pathogens, several comparable genomes were required for phylogenetics and pan-genomics studies. Initially, separate within-genera analyses were performed for each pathogen, Gram-positive L. monocytogenes and Gram-negative S. enterica. A comparison within the same genus, as in Listeria or Salmonella, was deemed the best approach for highlighting the genes contributing to human infection. The genera of Listeria and Salmonella vary greatly, and selecting non-pathogenic organisms within each genus was not possible. The non-pathogenic bacteria L. innocua Clip11262 and L. fleischmannii 1991 was selected for comparison to L. monocytogenes. The reptile pathogen S. bungori NCTC 12419 was the best genome with an alternative niche within the genus Salmonella. Model genome outgroups were selected as controls. The model Gram-positive bacterium Bacillus subtilis was selected as an outgroup for the Listeriae analysis. Two Gram-negative model pathogens, E. coli and Yersinia pestis, were chosen as outgroups for the Salmonellae probe. Y. pestis was included as an additional outgroup due to its highly virulent nature.

To provide perspective to pan-genomic studies and to verify evolutionary relationships of the organisms, phylogenetic trees were constructed utilising two methods. The first involved 16S rRNA genes as the basis for tree construction (Case et al., 2007). A multiple alignment (MUSCLE (Edgar, 2004)) was created from each consensus sequence (consambig (Rice et al., 2000)) of the 16S rRNA genes in each genome (RNAmmer (Lagesen et al., 2007)). A second approach aimed at creation of phylogenetic trees from the alignment of all the core genes in the pan-genome. This multiple alignment was created using the Roary tool (Page et al., 2015). The multiple alignments were analysed via RAxML phylogenetic inference software with 1000 replicate bootstrapping support (Stamatakis, 2014). The Newick format phylogenetic trees were visualised using Dendroscope (Huson and Scornavacca, 2012).

The Roary pan-genome analysis tool (Page et al., 2015) accepts annotated genome files from Prokka (Seemann, 2014) as input. To ensure that the gene function was conserved between the discovered orthologs, a higher BLASTP identity was preferred where possible. Separate pan-genome iterations were conducted for each set of genomes at BLASTP identities of 50, 70 and 90. The three different identities facilitated comparisons between the level of conservation in the Gram-positive or Gram-negative pan-genomes.

The annotated antibiotic resistance genes (McArthur et al., 2013) were analysed in the L. monocytogenes and S. enterica pathogens, and compared with related species. The resistance profiles of the non-pathogenic L. innocua and Gram-positive model B. subtilis were compared with that of L. monocytogenes. The cohort of resistance genes of the reptile pathogen S. bungori and Gram-negative model E. coli were contrasted with that of S. enterica.

3 Results and Discussion

The current research had a dual focus: to develop a pathogen biomarker identifier pipeline and to analyse the pathogenicity of sepsis pathogens. To this end, the concepts of a software pipeline were established and are illustrated in Figure 1. The overall pipeline was divided into two separate sub-pipelines. Sub-pipeline 1 performed the genome assembly and annotation based on a similar system with NSilico's Simplicity platform (Walsh et al., 2013). Sub-pipeline 2 contained the comparative genomics tools for the identification of pathogen biomarkers. The raw sequencing data assembled and annotated in Sub-pipeline 1 was compared with well-sequenced genomes derived from non-pathogenic organisms in Sub-pipeline 2.

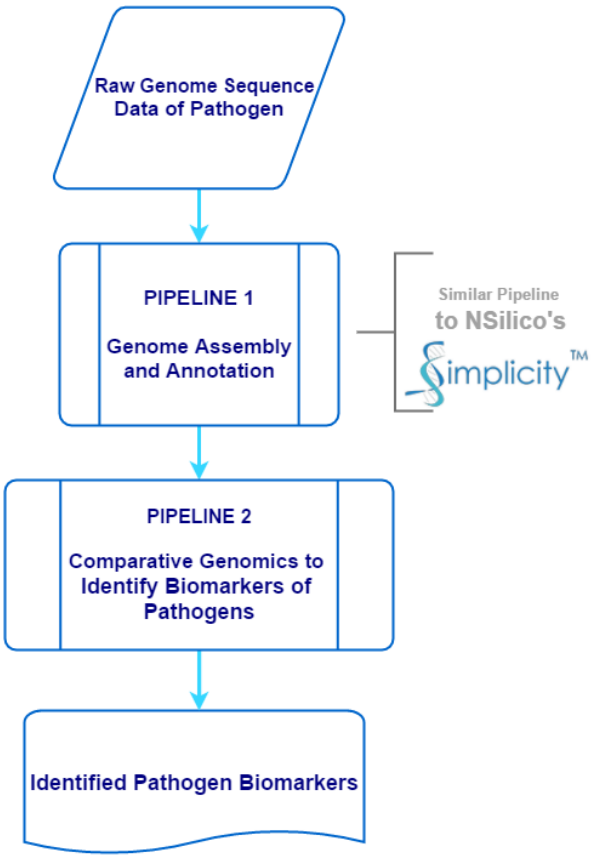


Figure 1: Workflow of the overall pipeline for the identification of pathogen biomarkers developed in this research.

3.1 Sub-pipeline 1 – Genome Assembly and Annotation

To select the most appropriate de novo assembler for our pipeline, two most frequently used tools, SPAdes (Bankevich et al., 2012) and Velvet (Zerbino, 2010) were examined using raw genomic sequences of L. monocytogenes and S. enterica. SPAdes was found to be superior than Velvet. The size of the L. monocytogenes genome is 2.94 Mbp (Glaser et al., 2001). Velvet assembled the L. monocytogenes genome in less but a sizable largest contig of 2.32 Mbp. In contrast, the largest contig yielded by SPAdes, for L. monocytogenes, reached 1.49 Mbp. Longer and fewer contigs are indicative of a good assembly as there are less gaps between contig ends, reducing the probability of error. The NGA50 statistic is given in base pairs and is calculated based on correctly aligned assembly sequence to the reference genome. The NGA50 is defined as, by aligning continuous blocks of the NGA50 length or longer, the blocks represent at least 50% of the reference genome

(Gurevich et al., 2013). The NGA50 of both assemblers was 1.49 Mbp for the L. monocytogenes genome. A misassembly in the largest contig assembled by Velvet caused over 800,000 bp to be incorrectly placed in the contig. Therefore, despite Velvet initially appearing to have constructed the best contig, both assemblers had the same maximum length of continuous correctly assembled sequence.

The number of complete genes in comparison to the reference genome was the most important assembly statistic, as genes were the focus of further downstream analyses. A lower number of complete genes is indicative of crucial areas of the genome being absent from the assembly. Velvet and SPAdes were compared in terms of their ability to assign higher number of complete genes to the given genomes. SPAdes was found to be superior than Velvet. SPAdes assigned 19 additional genes to L. monocytogenes genome, at a total of 3029. With regards to the S. enterica assembly using SPAdes, the number of complete genes was 108 genes greater than that produced by Velvet, at a total of 4724. Due to the higher number of complete genes, the SPAdes assembly was chosen here for both S. enterica and L. monocytogenes. SPAdes represents the first step in the Sub-pipeline 1.

The next step in the pipeline was to improve the de novo assemblies of the pathogen genomes, using the reference-assisted genome assembly approach. The reference-assisted assemblers including ABACAS (Assefa et al., 2009), CONTIGuator (Galardini et al., 2011) and Ragout (Kolmogorov et al., 2014) were compared using QUAST (Table 2) (Gurevich et al., 2013). ABACAS and CONTIGuator produced a single genome-sized scaffold out of the contigs for both the L. monocytogenes and S. enterica assemblies. However, the NGA50 statistics of ABACAS were far superior to that of CONTIGuator (Table 2). This was due to CONTIGuator introducing misassemblies not present in the ABACAS assembly. ABACAS also had a greater number of complete genes compared to CONTIGuator. Therefore, if a single scaffold is a desired goal, perhaps for studying alignment between genomes, ABACAS would be more useful than SPAdes.

After more detailed analysis, Ragout emerged as the preferred reference-assisted assembler. Ragout's method of using multiple reference genomes proved to be excellent at improving de novo assemblies. Using Ragout, the number of complete genes of the de novo SPAdes assembly, of both L. monocytogenes and S. enterica, increased dramatically (Table 2). Ragout also output the largest genomes, and greatly reduced the number of contigs. However, while Ragout was successful at extending and reducing the number of contigs, it did not merge all the contigs into a single scaffold like ABACAS. However, the higher number of complete genes outweighs one genome-sized scaffold. To improve upon the SPAdes de novo assembly, Ragout is designated the second step in the pipeline. The best assemblies were then annotated at the last step of Sub-pipeline 1.

3.2 Sub-pipeline 2 – Comparative Genomics

For comparative genomic studies, phylogenetic analysis was conducted employing two approaches of multiple alignment, 16S rRNA genes and the core genes from a pan-genome analysis. The cladograms that were created are depicted in Figure 2. Both sets of core genome alignments were the basis for excellent phylogenetic trees with bootstrap values of 100 for all branches (Figure 2B and D). The high bootstrap values are a stark contrast with the 16S rRNA based phylogenetic trees (Figure 2A and C). The branching order is incorrect in Figure 2A, with the correct order seen in Figure 2B. This erroneous ordering of the branches is an unfavourable measure of the 16S rRNA multiple alignment approach. The trees indicate that a multiple alignment of pan-genome core genes is superior for inferring a correct phylogenetic tree. One downside of the use of the core genome alignment, though, may be that it required far more computational power. The coupling of

phylogenetics and pan-genomics in the pipeline, using Roary (Page et al., 2015) and RAxML (Stamatakis, 2014), increases our understanding of the pathogens evolution.

	L. monocytogenes assembly				*S. enterica* assembly			
Statistic	SPA[1]	ABA[2]	CON[3]	RAG[4]	SPA	ABA	CON	RAG
Without reference								
No. of contigs	8	**1**	**1**	3	70	**1**	**1**	18
Largest contig (Mbp)	1.495	**2.947**	2.922	2.026	0.537	**4.876**	4.794	4.850
Total length (Mbp)	2.926	2.947	2.922	**2.948**	4.909	4.876	4.794	**4.961**
With reference								
No. of misassemblies	**1**	**1**	6	**1**	**0**	**0**	17	2
No. of N's per 100 kbp	**0**	688.1	23.95	15.26	**0**	1453.3	97.53	654.18
Genome fraction (%)	99.271	99.29	99.107	**99.943**	98.592	96.733	96.451	**99.025**
No. of complete genes	3029	3031	3028	**3042**	4724	4612	4603	**4768**
NGA50 (Mbp)	1.495	**2.885**	1.495	1.984	0.271	**4.785**	0.537	3.038

[1]SPAdes.

[2]ABACAS.

[3]CONTIGuator.

[4]Ragout.

Table 2: The comparison between the reference-assisted ABACAS (Assefa et al., 2009), CONTIGuator (Galardini et al., 2011) and Ragout (Kolmogorov et al., 2014) assemblies with the de novo SPAdes assembly (Bankevich et al., 2012). This comparison was performed through the QUAST tool (Gurevich et al., 2013). The statistics marked with and without reference indicate statistics that are measured against the reference genome. The statistics highlighted in bold indicate the best performance by that assembler for that category for that genome.

The antibiotic resistance of several related organisms was studied to identify the resistance profiles of sepsis pathogens. A set of Gram-positive bacteria were compared, which involved L. monocytogenes, L. innocua and B. subtilis. Another set of Gram-negative bacteria were contrasted, which included S. enterica, S. bungori and E. coli. A clear distinction was observed between the selected Gram-positive and Gram-negative collections of antibiotic resistance genes. A broad uniformity, in the numbers of antibiotic resistance genes, was observed within each group, Gram-positive and Gram-negative. Approximately 23 resistance genes were found in each of the Gram-positive bacteria, compared to 55 identified in the Gram-negative group. Evidently, antibiotic resistance was much more prevalent in the Gram-negative group. For example, genes for resistance to polymyxin and trimethoprim antibiotics were completely absent from the Gram-positive genomes. The most popular antibiotic resistance mechanism, across all genomes analysed, was the expulsion of antibiotic proteins from the cell (Wright, 2007). Alternative means of resisting antibiotics include enzymes, which break down the antibiotics, and proteins that replace the normal target of the antibiotic. The correct course of antibiotic treatment, for a neonatal sepsis patient, can

be determined by the antibiotic resistance profile of a pathogen (Reinhart et al., 2012). The huge extent of antibiotic resistance in Gram-negative pathogens suggests that novel drug targets are required urgently, while selecting appropriate antibiotics may be the best approach for Gram-positive pathogens.

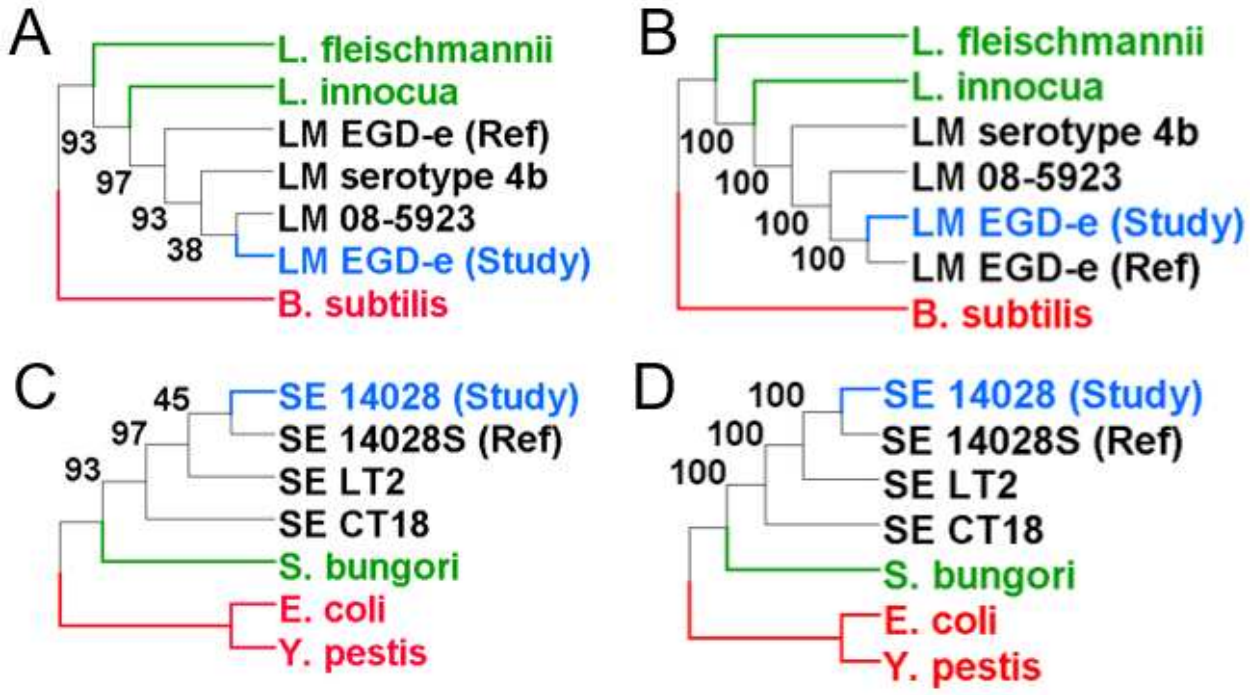


Figure 2. Cladograms generated using multiple alignments of 16S rRNA genes (A and C) or the core genome (B and D) for inferring the trees. A and B represent organisms from the Listeria genus, including B. subtilis as an outgroup. C and D represent organisms from the Salmonella genus, including E. coli and Y. pestis as outgroups. The blue nodes represent the pathogen under study. The green nodes are the non-pathogenic organisms, or non-human pathogens. The red nodes constitute the outgroups. The bootstrapping support is displayed for the branching order. LM, Listeria monocytogenes; SE, Salmonella enterica Typhi.

The use of the Roary tool (Page et al., 2015) for pan-genome analyses was effective as it constructed the pan-genome by using sequence comparisons. Using Roary, two different pan-genomes were created employing the organisms listed as nodes in either Figure 2A or Figure 2C. A clear disparity was observed in the level of conservation between the two groups. At a 90% level of identity, 6 common genes were found to be conserved in the core genome of the Gram-positive pan-genome. Five of this group of genes were annotated as ribosomal proteins, additionally to a single stress tolerance protein. Adjusting the percentage identity cut-off to 50%, inflated the core genome to 850 genes. At a 90% identity cut-off, 188 common genes were found in the Gram-negative core genome. Like the Gram-positive group, many (almost 40) conserved ribosomal proteins were among this 188 gene set, in addition to the genes linked with increased motility of Gram-negative bacteria. Over 2000 genes were conserved in the Gram-negative core genome at a 50% identity threshold. This difference in conservation of core genes reflects different levels of evolution and diversity in Gram-positive (L. monocytogenes) and Gram-negative (S. enterica) genomes. The lower number of conserved genes in the Gram-positive genomes could indicate their coevolution and dependence on host genomes (Glaser et al., 2001). Opposing this, the Gram-negative genomes

demonstrate genetic flexibility reflected by their ability to survive in diverse ecological niches and cause different types of infections involving for example, S. enterica Typhi organisms (McClelland et al., 2001). The ribosomal proteins that are highly conserved across the genomes could serve as attractive drug targets to attack the pathogen.

To gain a better understanding of the sepsis pathogens, subsets of the pan-genomes were analysed in L. monocytogenes and S. enterica. Genes present in L. monocytogenes EGD-e, but absent from non-pathogenic L. innocua were identified using comparative genomics approach. Seven members of the internalin coding gene family were found among the unique group of genes in L. monocytogenes. The internalin proteins allow the L. monocytogenes bacterium to enter mammalian cells (Glaser et al., 2001). The S. enterica Typhi 14028 pathogen genome was compared with that of S. bungori. Several genes that were found to be unique to S. enterica Typhi, code for invasion related proteins such as SopB and SipC, enabling S. enterica Typhi to invade human cells (McClelland et al., 2001). The identification of the virulence proteins mentioned above, validates the functionality of the pathogen biomarker pipeline. Many putative proteins were found to be unique to the pathogens. These unexplored proteins may be a useful starting point to investigate the pathogen in more detail.

4 Conclusions and Future Work

Comparative genomics is a growing field with a diverse range of possible endeavours (Holden et al., 2013; Vernikos et al., 2015). Some genome comparative methods were applied in this research. The next step would be to study and compare the functional groups of orthologous genes among bacteria (Brettin et al., 2015). Pan-genomics analysis represents the most promising aspect of this pipeline. The permutations incorporating different bacteria, achieved via pan-genomics analysis, are infinitely numerous. Subsets of genes obtained from pan-genomics, such as genes unique to a genome or core genes, may be analysed for their role in metabolic pathways. Essential genes coding for transcription factors, that regulate the protein transcription process, could be targeted via next generation drugs. This in silico information serves as an important basis for experimental validation of the nature of a pathogen. A future pursuit for this comparative genomic pipeline is to develop a system allowing antibiotic resistance gene identification of all bacterial strains implicated in neonatal sepsis infection.

To our best knowledge, there is no broad-spectrum tool for comparative genomics. Therefore, there is a huge potential to progress further with this pipeline. Simplicity (Walsh et al., 2013) includes a comparative genomics pipeline that will be further developed by application of the knowledge accumulated during this research. An aim of this research was to develop an automated pathogen biomarker identification tool which would aid researchers and clinicians with limited bioinformatics expertise. The in silico comparative analyses of the pipeline have the capacity to support and accelerate the research of bacterial pathogens. This pipeline allows the clinician to upload sequenced pathogen samples for initial de novo assembly and annotation, and for subsequent comparison and antibiotic resistance biomarker identification in vitro.

5 References

Alexander DC, Fitzgerald SF, DePaulo R, Kitzul R, Daku D, Levett PN, Cameron ADS (2016) Laboratory-Acquired Infection with Salmonella enterica Serovar Typhimurium Exposed by Whole-Genome Sequencing Diekema DJ, ed. J Clin Microbiol, 54, pp. 190-193.

Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS (2016) PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res, 44, pp. W16–W21.

Assefa S, Keane TM, Otto TD, Newbold C, Berriman M (2009) ABACAS: Algorithm-based automatic contiguation of assembled sequences. Bioinformatics, 25, pp. 1968–1969.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V, Sirotkin A V, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol, 19, pp. 455–477.

Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason III JA, Stevens R, Vonstein V, Wattam AR, Xia F (2015) RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep, 5.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics, 10, pp. 1–9.

Case RJ, Boucher Y, Dahllof I, Holmstrom C, Doolittle WF, Kjelleberg S (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol, 73, pp. 278–288.

Chen L, Zheng D, Liu B, Yang J, Jin Q (2016) VFDB 2016: Hierarchical and refined dataset for big data analysis - 10 years on. Nucleic Acids Res, 44, pp. D694–D697.

Dickinson P, Smith CL, Forster T, Craigon M, Ross AJ, Khondoker MR, Ivens A, Lynn DJ, Orme J, Jackson A, Lacaze P, Flanagan KL, Stenson BJ, Ghazal P (2015) Whole blood gene expression profiling of neonates with confirmed bacterial sepsis. Genomics Data, 3, pp. 41–48.

Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol, 7, pp. e1002195.

Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, 32, pp. 1792–1797.

Ekblom R, Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl, 7, pp. 1026–1042.

Fortier L-C, Sekulovic O (2013) Importance of prophages to evolution and virulence of bacterial pathogens. Virulence, 4, pp. 354–365.

Galardini M, Biondi EG, Bazzicalupo M, Mengoni A (2011) CONTIGuator: A bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol Med, 6.

Glaser P et al. (2001) Comparative genomics of Listeria species. Science (80-), 294, pp. 849–852.

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: Quality assessment tool for genome assemblies. Bioinformatics, 29, pp. 1072–1075.

Holden MTG et al. (2013) A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. Genome Res, 23, pp. 653–664.

Huson DH, Scornavacca C (2012) Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. Syst Biol.

Illumina (2014) De Novo Assembly of Bacterial Genomes. Illumina Appl Note, pp. 5–8.

Kolmogorov M, Raney B, Paten B, Pham S (2014) Ragout - A reference-assisted assembly tool for bacterial genomes. Bioinformatics, 30, pp. I302–I309.

Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW (2007) RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res, 35, pp. 3100–3108.

Liu L, Oza S, Hogan D, Perin J, Rudan I, Lawn JE, Cousens S, Mathers C, Black RE (2015) Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: An updated systematic analysis. Lancet, 385, pp. 430–440.

McArthur AG et al. (2013) The Comprehensive Antibiotic Resistance Database. Antimicrob Agents Chemother, 57, pp. 3348–3357.

McClelland M et al. (2001) Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. Nature, 413, pp. 852–856.

Niu C, Yu D, Wang Y, Ren H, Jin Y, Zhou W, Li B, Cheng Y, Yue J, Gao Z, Liang L (2013) Common and pathogen-specific virulence factors are different in function and structure. Virulence, 4, pp. 473–482.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J (2015) Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics, 31, pp. 3691–3693.

Rathinam VA, Vanaja SK, Fitzgerald KA (2012) Regulation of inflammasome signaling. Nat Immunol, 13, pp. 332–333.

Reinhart K, Bauer M, Riedemann NC, Hartog CS (2012) New approaches to sepsis: Molecular diagnostics and biomarkers. Clin Microbiol Rev, 25, pp. 609–634.

Rice P, Longden L, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet, 16, pp. 276–277.

Seemann T (2014) Prokka: Rapid prokaryotic genome annotation. Bioinformatics, 30, pp. 2068–2069.

Simonsen KA, Anderson-Berry AL, Delair SF, Dele Davies H (2014) Early-onset neonatal sepsis. Clin Microbiol Rev, 27, pp. 21–47.

Smith CL, Dickinson P, Forster T, Craigon M, Ross A, Khondoker MR, France R, Ivens A, Lynn DJ, Orme J, Jackson A, Lacaze P, Flanagan KL, Stenson BJ, Ghazal P (2014) Identification of a human neonatal immune-metabolic network associated with bacterial infection. Nat Commun, 5.

Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30, pp. 1312–1313.

Stoesser N, Moore CE, Pocock JM, An KP, Emary K, Carter M, Sona S, Poda S, Day N, Kumar V, Parry CM (2013) Pediatric bloodstream infections in Cambodia, 2007 to 2011. Pediatr Infect Dis J 32:e272–e276.

Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L (2015) Update on RefSeq microbial genomes resources. Nucleic Acids Res 43:D599–D605.

Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154.

Walsh P, Carroll J, Sleator RD (2013) Accelerating in silico research with workflows: A lesson in Simplicity. Comput Biol Med 43:2028–2035.

Wright GD (2007) The antibiotic resistome: The nexus of chemical and genetic diversity. Nat Rev Microbiol 5:175–186.

Zerbino DR (2010) Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinforma:11.5.1-11.5.12.

An approach for anonymization of sensitive clinical and genetic data based on Data Cube Structures

Ntalaperas Dimitrios, Bouras Athanassios
Ubitech Ltd, Greece
dntalaperas@ubitech.eu

Keywords

Anonymization, Security, Algorithms

Abstract

This paper presents an approach for anonymizing data containing sensitive patient information. The approach is based on the usage of a Data Cube structure that stores only aggregate information of the various phenotypic and genetic values of variables of interest. The Data Cube is further perturbed in a manner that does not affect the statistical characteristics of the original dataset, thus further enhancing the fidelity of the anonymization procedure. The anonymized data can be semantically enhanced and transferred over the network in plain format without the risk of violating patients' personal rights.

1 Introduction

As Electronic Health Records (EHR) are adopted by an ever growing number of clinical health care providers, institutes can easily interchange information electronically, which can then be statistically analyzed, thus increasing research output. However, interchange of data concerning clinical trials and genetic data are subject to protection laws, which inhibit the disclosure of sensitive patient information. To this extent, data should be anonymized, not only to ensure that sensitive data is not eavesdropped during transmission, but also to ensure that the party legitimately receiving the data cannot have access to sensitive personal data.

In this paper, we present a methodology for transforming data corresponding to patient clinical and genetic data to an equivalent data set that contains the informational content of the original data set regarding the medical data, but from which all information regarding personal data has been filtered out. The methodology builds upon the Data Cube approach as this has been adopted by the Linked2Safety [1] consortium (Perakis et al. 2013) combined with cell-suppression and perturbation techniques which ensure that the produced data cube cannot be reversed engineered (Forgó et al. 2012), while also retaining the same statistical characteristics as the original data set.

The resulting methodology is being used in the context of the SAGE-CARE [2] project, in order to be able to transmit data that combine phenotypical characteristics and counts of gene expressions of a patient in a safe manner, retaining patients' anonymity while at the same time allowing researchers to search for correlations between these data in the context of melanoma research.

[1] http://www.linked2safety-project.eu
[2] http://www.sage-care.eu

2 Theory

A dataset containing medical data can have mixed information with sensitive data being present at various locations of the source files. Moreover, the informational content may be in semi-structured or unstructured format (e.g. free text). An anonymized dataset must fulfil the following criteria:

1. Any information of personal data must be filtered out
2. Information that can be used to reverse engineer the anonymized data must be filtered out. If, for example, a combination of phenotypic characteristics is extremely rare and this data is present in the anonymized set, an eavesdropper can use this information to identify a patient.
3. Any statistical analysis performed on the anonymized data should yield the same results as for the original set.

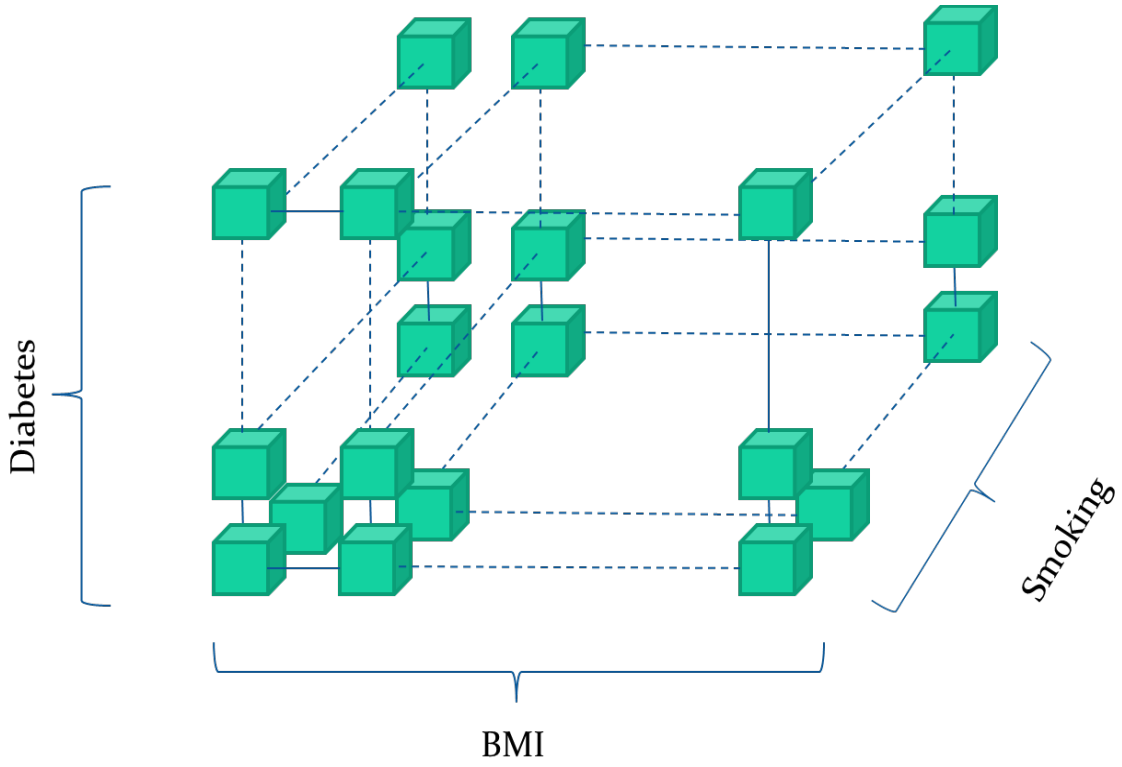


Figure 1. Example of a Data Cube Structure.

A data structure that can fulfil the above criteria is a Data Cube. A Data Cube can be defined as an $m_1 x m_2 x ... x m_n$ array, with n being the number of selected variables and m_i being the number of distinct values the variable indexed by i can take. Each cell contains the number of total counts with the combination of values defined by the index of the Data Cube. Figure 1 depicts an example of a three-dimensional Data Cube. In this example, each variable is supposed to be categorical (take values from a district set of integers). The value *0* for BMI for example can correspond to low BMI, value *0* for diabetes to low blood sugar levels and so on. If the Data Cube is named *DC*, then the cell *DC[i][j][k]*, will have the total count of patients having BMI equal to *i*, Smoking equal to *j* and Diabetes equal to *k*. Since the cube contain only aggregate data, it fulfils criterion 1. Moreover, all data correlations of the original dataset concerning these variables are not altered, thus criterion 2 is satisfied.

A special case is that of *degenerate* data cubes, that can arise from non-categorical data. If a variable is not categorized (e.g. has continuous values) and a data cube is constructed naively, then it may be the case that each distinct value will correspond to a very small number of patients. In this case, the data cube will have a size of the order of the size of the original data set with each containing a small number. Criterion 2 in this case will be violated.

The case of degenerate cells is handled by adopting a combination of cell-suppression and perturbation techniques as well as imposing categorization on the data. These two techniques are described in the following sub-sections.

2.1 Cell-suppression

Cell-suppression is the procedure that forces cells with a low value to be discarded from the dataset. By discarding these cells, the process of distinguishing non-existent combinations from combinations with a low count becomes impossible. Perturbation imposes an extra layer of security by imposing a random, low amplitude noise to each cell so they are no longer guaranteed to contain the exact number of patients sharing the characteristics described by the cell. The distribution of the noise can always be chosen as to not alter the statistical characteristics of the original dataset. A typical configuration is to randomly select a value from the set

{-1,0,1} to add to the final count with equal probability. This combination ensures in most cases that the statistical characteristics of the anonymized dataset will remain unaltered (Antoniades et al. 2012) thus ensuring that Criterion 3 is still being fulfilled.

Considering the specific distribution that occur in each data set, the user may opt to increase the perturbation noise to further obfuscate the data or to decrease it to keep the fidelity of the data as high as possible. In general, as the ratio of distinct combinations to total patient grows, so does the range of perturbation noise.

2.2 Data Categorization

Data Categorization refers to the process of mapping values of source data, that take continuous or arbitrary values, to values of a finite set. The mapping does not need to be isomorphic, in fact, in the case of continuous mapping, isomorphic mapping is impossible. Figure 2 depicts an example for each of the cases of Data Categorization.

In the case of BMI the source data contain entries that describe the BMI of each patient and the categorization is performed according to the international classification (WHO 1995). This is often the case of numerical data, since these usually correspond to phenotypic variables that can be categorized according to medical standards or according to various taxonomies and ontologies like MedRA [3] or SEHR (Sahay et al. 2013). In the case of Smoking on the other hand, the values are arbitrary and can be numerical or text. The mapping in this case must be custom and be provided explicitly via a mapping file. A mapping file is also needed in the case of numerical data, if the user does not wish to conform to a standard ontology.

In the context of SAGE-CARE, both methods have been implemented. Phenotypic variables are typically mapped by using standard taxonomies, whereas genetic data (typically counts of gene expressions) are mapped in a custom way, since the needs of categorization depend both on the specific gene, as well as the context of the specific correlations the user wishes to derive.

[3] http://www.meddra.org

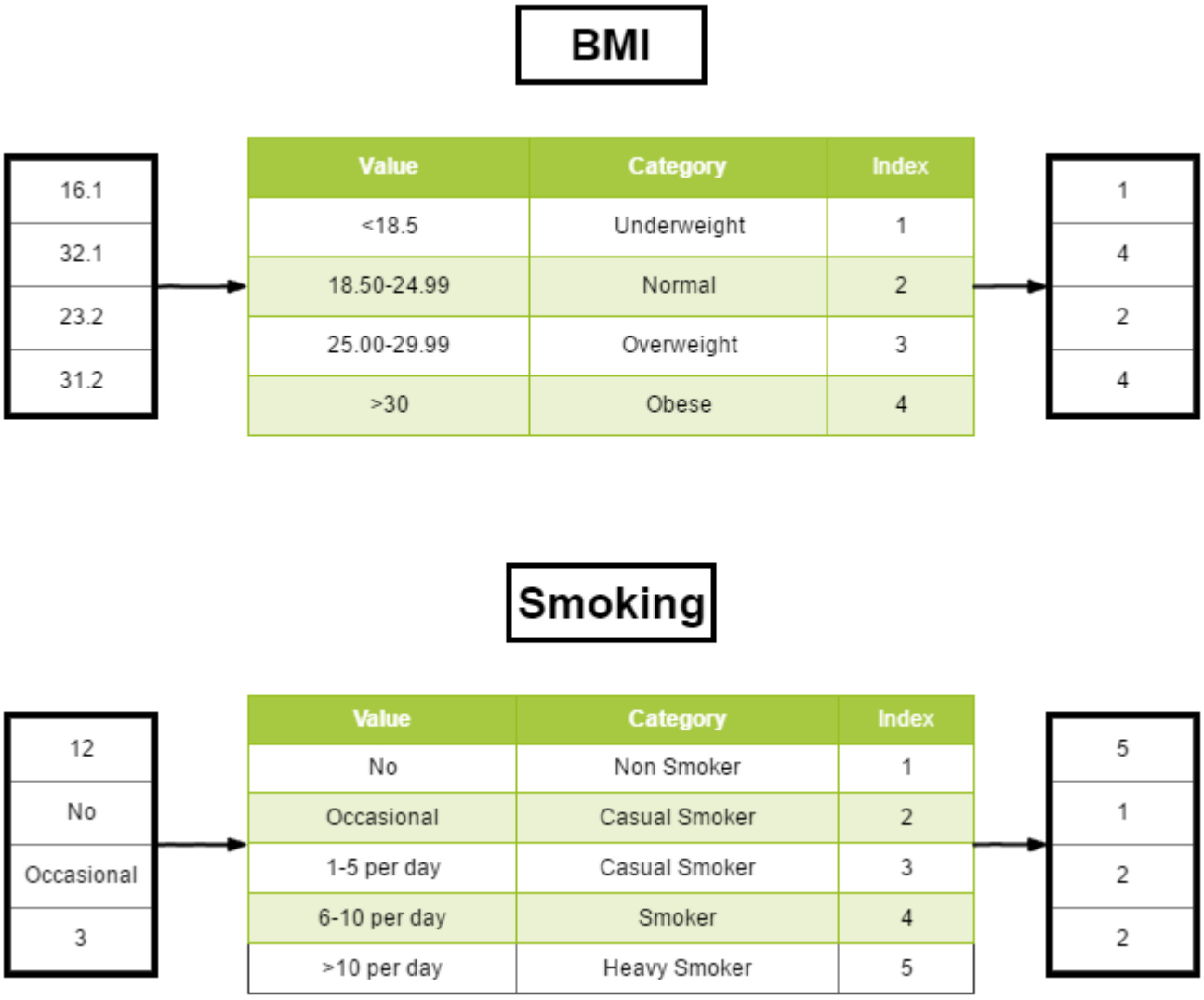


Figure 2. Categorization Examples

3 Implementation

The above general techniques presented above were implemented in the context of the SAGE-CARE project to create anonymized clinical (phenotypical) data and genetic data recording the expression of genes of each patient. The methodology pipeline is depicted in Figure 4. The component at first accepts the two files that contain the source data; a .csv file that contains phenotypical data with each row corresponding to each patient and a .tsv file that contains genetic data with each row corresponding to a gene and each column to a patient. The .tsv file contains the patient code at the top cell of each column thus facilitating the linking of patients between the two datasets. The component constructs an intermediate structure which aligns the two data set so that operations, such as aggregation which is required for data cubes, can be easily performed. An example depicting how data alignment is performed can be seen in Figure 3, where the phenotypic variables of a fake patient is matched against the file containing genetic information. As can be seen the barcode of the patient is used to obtain the column, indexed *j*, of the genetic data file. Supposing that we need to construct a cube containing expression counts of the gene with Entrez id equal to *87769*, the corresponding row, indexed *i* of the *.tsv* file needs to be found by matching the id against

the values of the cells of the first column. The cell with the coordinates *(i,j)* of the *.tsv* file contains the value for the expression count that is to be stored to the data aligned data structure. The header of the column, which is to be used to identify the genetic variable, will be the Entrez id of the gene.

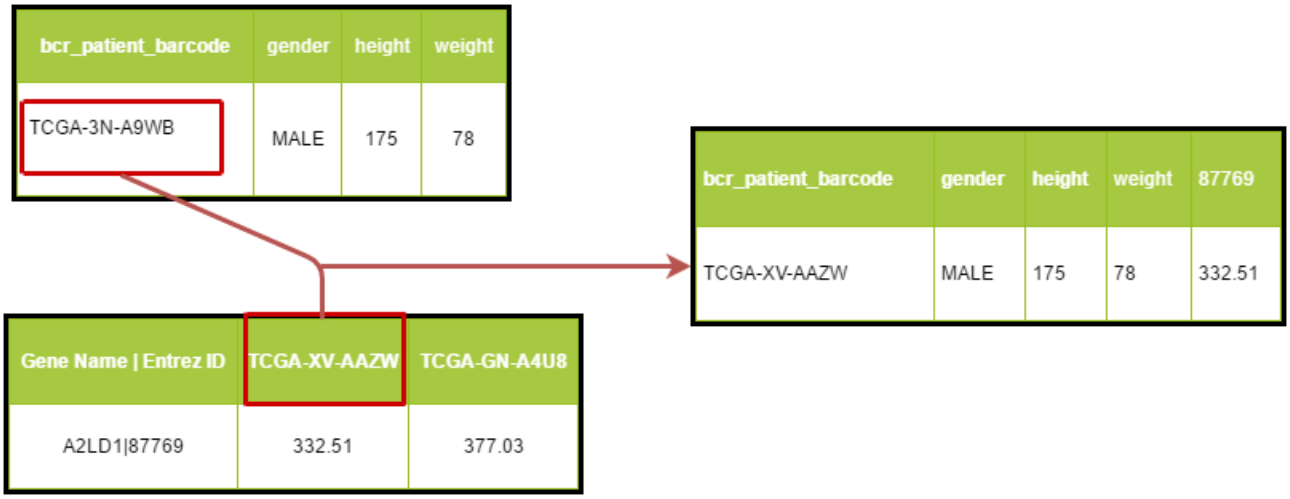


Figure 3. Data Alignment example for one patient.

At the second state, the component receives a mapping file in .xml format which contains the variables and genes for which a data cube is to be created. The mapping file also contains information for grouping values to categories; the categorization can be performed either by using value ranges or by using enumerations (see Listing 1 for a sample mapping file). Alternatively, as discussed in Section 2, a taxonomy can be used.

The data are transformed according to the categorization schema and are then fed as input to the Data Cube Creation component. The Data Cube Creation component performs aggregation on the data (see Listing 2 for pseudocode describing the algorithm).

Having the Categorization being performed after the Data Alignment can have a detrimental effect on performance. This can easily be seen from the fact that the tables containing the aligned data need to be parsed again in order to convert original values to categorical ones, something that could have been done during Data Alignment process. Keeping however the original (uncategorized) data separate, has the advantage that these can be reused with different categorization schemas, something that reduces execution times of subsequent calls to the component.

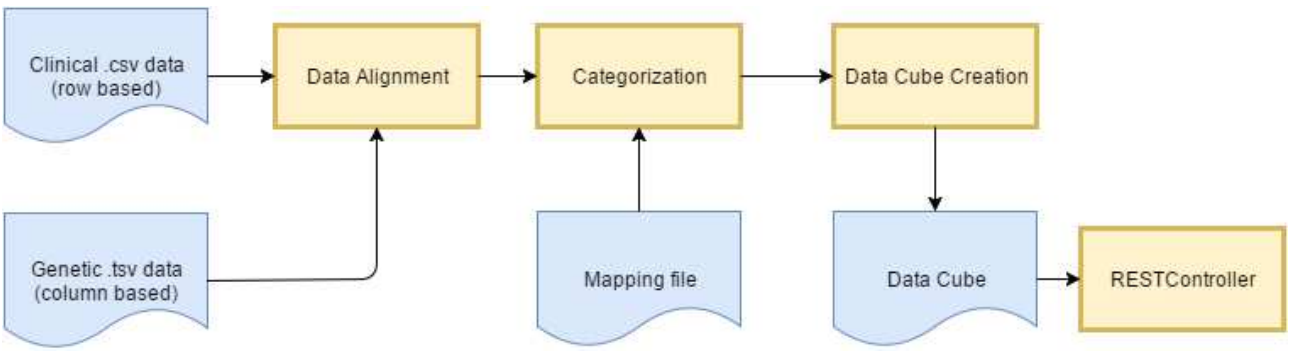


Figure 4. Methodology Pipeline

```
<?xml version="1.0" encoding="UTF-8"?>

<root>

    <phenos>

        <pheno name="gender" />

        <pheno name="height" round="1d" />

    </phenos>

    <genes>

        <gene name="155060" round="3d" />

    </genes>

</root>
```

Listing 1. Sample mapping file. Variable gender is defined without a mapping, which implies that the data are already categorized regarding gender values. Variable height is to be rounded to first digit which means that values within 10 cms will be grouped together. Similar for the gene with entrez code 155060 the expression counts within 1000 will be grouped together.

```
procedure dataCube
        foreach line
                computeunique_key
                foreach cell in line_cells
                        cell_field++
                        value[cell_field] = valueOf(cell)
                        value[cell_field] = convertInRange(value, ranges[cell_field])
                Cube[unique_key][indexOf(value[cell_field])]++
        Foreach unique_key
                for i=1 to num_of_fields
                        perturbe(Cube[unique_key][i])
                        if Cube[unique_key][i] <= threshold
                                discard (Cube[unique_key][i])
```

Listing 2 Data Cube Creation Algorithm. Unique key is obtained by computing a unique hash from the combination of variables values corresponding to each cell.

3.1 Special Case: Genetic Data

It is often the case that sets of phenotypic variables need to be checked for correlations against a large dataset of genetic data containing thousands of genes or DNA base pairs. Constructing a Data Cube for all these genes may cause degeneracy, long execution times and will require a vast

amounts of storage space, considering that the Data Cubes will have a dimensionality on the order of thousands.

However, obtaining correlations between thousands of genes is seldom needed in practice; often, what the user needs is deriving correlations between a set of phenotypical variables and a gene, or between a small number of genes. The latter case can be handled by the techniques already discussed, for the former case the Data Cube component can be executed in a special mode. The algorithm of this mode is depicted in Listing 3. The procedure is simple: for each gene, a data cube is created using the set of phenotypic variables plus the gene. If the number of phenotypic variables is *n* and the number of genes is *m*, this procedure will create *m* data cubes of dimensionality *n+1* (*n* being the phenotypic variables plus one for the gene variable). Using the naïve approach a single data cube would be created with a dimensionality equal to *m+n*.

```
procedure geneticDataMode
        foreach gene
        call dataCubeGene
```

Listing 3 Data Cube Creation Algorithm for Genetic data

4 Conclusions

In this paper a methodology for ensuring the anonymization of data present in the context of the SAGE-CARE project was presented. The methodology is based on peer reviewed structures and techniques that are validated both technically and legally thus providing a framework for the required level of data security.

Future work consist of exploring ways to align data between heterogeneous data sources that are not covered by the mapping scheme presented here. Under consideration is the adoption of OpenRefine[4], which allows the filtering, faceting and noise extraction on data, as well as the definition of a common mapping ontology in a similar fashion that was followed by the Linked2Safety consortium.

Furthermore, there is ongoing research, with the collaboration of University of Naples [5], regarding ways to increase the performance of the algorithm by using High Performance Computing (HPC) techniques as to be able in the future to compute multidimensional data cubes and generate responses to queries in real time. The two points of the algorithms that are considered is the data alignment and the aggregation process; the tabular data that these two procedures are manipulating do not, in general, have interdependence and thus parallelization of these two routines can lead to significant gains in performance.

[4] http://openrefine.org

[5] http://www.unina.it

5 Acknowledgment

We would like to thank Dr. Paul Walsh for providing guidance about what the user expectations of the component could be and for providing information about the representation of original source data. We would also like to thank Professor Bernhard Humm for providing insight on how to best use ontology for data mapping and categorization. Last, but not least, we would like to thank Professor Davide Marocco for his help regarding the parallelization of the algorithm.

6 References

K.Perakis et. al. (2013). "Advancing Patient Record Safety and EHR Semantic Interoperability", in: IEEE International Conference on Systems, Man, and Cybernetics, IEE, 2013.

N. Forgó, M. Góralczyk, and C. Graf von Rex. (2012) "Security issues in research projects with patient's medical data", in: 12th Int. IEEE Conf. on Bioinformatics & Bioengineering (BIBE), IEEE, 2012

A. Antoniades et. al. (2012). "The effects of applying cell-suppression and perturbation to aggregated genetic data", in: 12th Int. IEEE Conf. on Bioinformatics & Bioengineering (BIBE), IEEE, 2012.

WHO (1995). Physical status: the use and interpretation of anthropometry. Report of a WHO Expert Committee. WHO Technical Report Series 854. Geneva: World Health Organization, 1995.

R. Sahay et. Al. (2013). "An Ontology for Clinical Trial Data Integration", in IEEE SMC 2013 - IEEE International Conference on Systems, Man, and Cybernetics, IEE, 2013

Chapter 6

Humanities / Social

Designing narrative artefacts – the lives of women in eighteenth century Cork

J. Dempsey, P. Green, A. Wilson
Department of Media Communications
CIT – Cork Institute of Technology, Ireland
jenny.dempsey@mycit.ie

Keywords

narrative, design, material culture, heritage

Abstract

This paper introduces the initial phase of a research project which is aligned with the development of a new heritage centre in eighteenth century Cork City, Ireland. The topical content of the research is focussed on the impact of the social and political environment at the time on the lives of women. The project uses narrative, integrated with visual design and material culture, to support the goal of producing engaging experiences for visitors and remote users of the heritage site. The project intends to contribute to the study of narrative in art and design by exploring how audiences participate in the recovery of stories from material objects. The paper provides a context for the project, an overview of the methods, and a description of some prototypes that are currently being developed for evaluation through the educational programme to be established by the centre.

1 Introduction

The research project introduced here is aligned with the development of a new heritage centre in Cork City entitled Nano Nagle Place which is due to open to the public in early 2017. The centre is dedicated to communicating the life story and legacy of Honora (Nano) Nagle. Part of this story involves presenting a portrait of the city of Cork against a backdrop of political and social life in eighteenth century Ireland. The topical content of the research is embedded within this history and focusses specifically on the impact of the social and political environment at the time on the lives of women.

Academically, the project investigates the way in which narrative, when integrated with visual design and material culture, can support the goal of producing engaging and memorable designs for visitors or remote users of the heritage site. The project therefore contributes to the emerging study of narrative in art and design (Anderson, 2011, Hadjilouca et al., 2013) by accounting for the way in which visual media can be employed to allow audiences participate in the recovery of stories from material objects.

2 Material Culture

Material culture involves the study of material things and the values and roles associated with them in different societies (Gerritsen and Riello, 2015, p.2). Objects and the physical environment have long been central to research in archaeology, anthropology, art history and museum studies, but modern historians only began to take an interest in them from about the 1960s, when an interest in ‘history from below’, or the history of ordinary people, became seen as increasingly worthy of study. The majority of human beings have left no trace in the archival texts conventionally used by historians, which are mainly written by, and refer to, those considered relatively powerful and noteworthy. Material culture came to be seen as a fruitful way of circumventing this problem, by

studying the objects with which people interacted, and which they used in a variety of ways both to sustain their lives and to communicate personal and social meanings and values (Dant, 1999, p.38).

As research into this area developed, the study of material objects along with other sources came increasingly to be seen as a way of gaining a greater understanding generally of the past, of what it felt like to have lived in a particular historical period. The inclusion of material artefacts as sources also prompted researchers to ask new questions. Objects are products, but also shapers, of human activity, needs and values, and consequently central to how and why humans interact in particular ways with each other and with their environment (Harvey, 2009, p.5). Studies of the production, consumption and international trade in objects have contributed significantly to our understanding of historical societies and how and why they developed in particular ways, while close examination of specific items, or groups of items, have also provided insights into individual and social lifestyles and values.

This project foregrounds objects (original artefacts and replicas, as well as those represented in texts and images) in order to enrich people's (and particularly children's) understanding of a particular historical period and location, eighteenth-century Cork. It focuses on the lives of women, a group which has been disproportionately poorly represented in traditional historical sources. An engagement with material things is used to make people think closely about daily activities and how they were managed in eighteenth-century Cork, but also to help them imagine the physical, sensory environment in which these activities took place, mobilizing as much as possible senses such as touch, smell and taste. The objects, in conjunction with other sources, also offer ways of gaining insight into economic, class and gender divisions of the period, as different items were available to, or used differently by, various groups. Some of the objects are unfamiliar and strange to contemporary audiences, while others still play a significant (if often different) role in modern life, but both can be used to enhance understanding of both the strangeness and familiarity of daily life for women in Cork over two hundred years ago.

3 Narratology and Objects

Narratology in the twentieth century has largely emerged through work done in the literary arts. Francophone structuralism in the 1960s popularised the notion of a science of storytelling. By adapting earlier studies in structural linguistics, structural anthropology, and formalism, morphology, or grammar of stories was imagined. The idea that the underlying structural rules of narrative could be exposed and catalogued was the initial force that While initially this project was presented as broad giving recognition to the ubiquity of stories in all aspects of human culture and life (Barthes, [1966] 1977) the narratological project was driven mainly by researchers in literary and language studies. This line of investigation led an enormous body of information on the structural, reader centric, rhetorical, cognitive aspects of fiction as it is represented in the history, mostly, of literature.

In parallel to the work achieved in literary narratology there are other fields which harboured and interest in narrative which attracted less attention. Within the social sciences there is a narrative genealogy which extends from William Labov (Labov, 1972) to present studies where narrative analysis has been popularised as a methodology through the work of Carol Riessman (1993). By the turn of the millennium the explosion of interest in narrative was felt across a range of disciplines influenced by a constructivist orientation in psychology (Bruner, 1986, 1990). Current theoretical discourses such as cognitive narratology (Ryan, 1991, Herman, 2003, 2010, Fludernik, 2003), transmedial narrative (Herman, 2004, 2012, Ryan, 2004, Jenkins, 2006) and multimodal narrative

(Page, 2010) have opened up possibilities for the appreciation of narrative in art and design (Anderson, 2011, Hadjilouca et al., 2013).

These theories are regularly cited in narrative approaches to architecture, art history, new media, and design. While cultural artefacts such as novels and films were once seen as the prototype media for storytelling, in recent years we can more easily find projects using material objects, such as maps (Speed, 2009, Flaherty, 2011), buildings and architectural spaces (Psarra, 2009, Coates, 2012, Macleod et al., 2012), or personal items (Speed, 2010), that are presented as things from which stories can be recovered. This research shares an interest with design projects that display an object-oriented approach to narrative.

4 Approach & Methodology

The project has access to various onsite facilities at Nano Nagle Place as indicated in Figure 1. These include a permanent exhibition space, a temporary exhibition space, an education room, rooms dedicated to the delivery of educational content and workshop activities, as well as outdoor spaces which are networked with high capacity digital communications. There will also be opportunities for the project to connect with audiences through the centre's outreach educational programme which will support and augment the curriculum in primary and secondary schools in the region. The research project is therefore related to learning but more explicitly it aims to engage the audience with eighteenth century historical information.

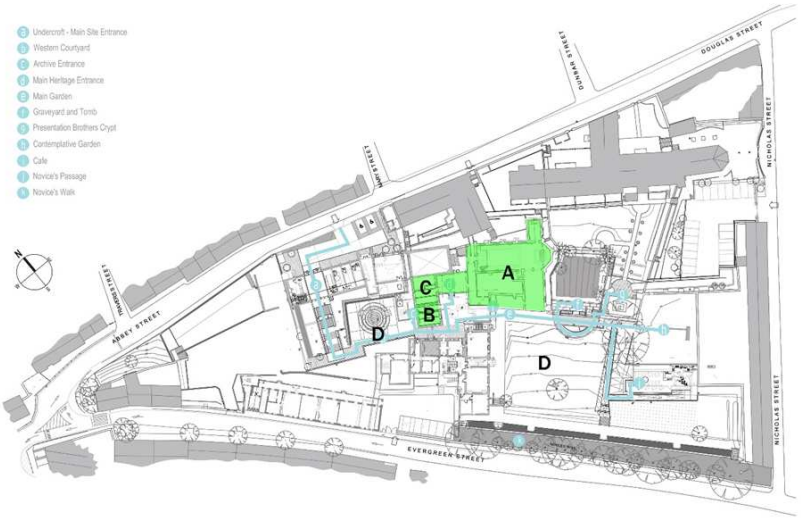


Figure 1. The site plan for Nano Nagle Place (approx. 3.8 acres) which includes potential target locations for the research onsite. The area marked A represents the location of the permanent exhibition; B is the proposed site of the temporary exhibition space; C is the education room and workshop spaces for young visitors; D represents potential outdoor spaces also available for use by the research project.

The development of the research was initially organised into four discreet phases: 1) data collection; 2) design investigations; 3) prototype/artefact development; and 4) the review and presentation of overall findings. Each stage involves its own cycle of investigation which results in a report of the findings related to that phase. While each stage can be regarded as discreet in terms of its objectives there is also significant overlap between the stages. For example, while stage one was intentionally dominated by activities such as: the review of historical material directly associated with Nano

Nagle and the Presentation Sisters; site visits to relevant archives and museums; and desk reviews of literature on narrative theory; there was also some visualisation exercises and paper prototyping of ideas as information was being collected and evaluated.

The first stage resulted in a body of information which provided a snap shot of eighteenth century Cork when Nano Nagle was establishing her schools for catholic children. This historical information helped provide a picture of the conditions under which women from different social and economic standing lived. This provided a basis for the development of narrative accounts which would be designed with a view of engaging visitors to the heritage centre, or young audiences participating in the educational programme.

Imagining how women of different social standing may have experienced Cork in the 1700s provided the basis for the development of storied content for prototype designs. Early stage paper prototypes mock-ups of narrative objects were presented in June 2016 to stakeholders who included members of Nano Nagle Place Heritage Committee, and the JCA architects responsible for the design and layout of the overall site and exhibition spaces.

Initial Prototypes

There were two initial prototypes presented for discussion with stakeholders. The first of these was a paper prototype of an interactive book, and the second explored picture cubes as a means of providing different narrative possibilities for the end user to explore. The prototypes were designed to initiate a dialogue with the stakeholders around a central objective which was generate visitor/young audience engagement with eighteenth century Cork history.

The first prototype, the interactive book, was intended to be located in the section of the permanent exhibition presenting Cork City as the mise en scène for the (hi)story of Nano Nagle's philanthropic work. Based on the historical data that was accumulated in stage one, five fictional female characters from different social standings were designed and a vignette of their lives was illustrated. One character was selected for further development as fictional content for the book, shown in Figure 2. The structure of the book proposed to take the user through a set of events associated with the purchase of material for the design of a dress. This operated as a conventional book with the plot unfolding as the user flips through the pages. This facilitated the story of the character's journey through the city offering the user an introduction to the 'historical' city by taking them to shops in specific streets in the city. The book however also promised to offer interactive options for accessing, and thereby directly experiencing, elements of the material culture of the time through the physical attributes of the book itself. The elements considered included the materials used in the making of the book, or physical elements that were part of the story but which could be removed and explored by the user. Examples included a piece of cloth fashionable and of the time, or patterned paper used for handwritten letters. Some consideration was also give as to how new media technologies could be integrated with the objects to provide further options for interaction and feedback.

Figure 2. Paper prototype of interactive book with embedded material elements. Considerations were given to how a narrative account could be facilitated by the materials themselves or extended by digitally augmenting the elements.

A second prototype, shown in Figure 3, explored the idea of designing blocks that would allow users to assemble images into narrative sequences. Only certain coded sequences would offer the user with the reward of a story which could be recovered in a number of ways. By being digitally augmented certain sequences could automatically activate an audio account or unlock further clues in a block encouraging the user to 'discover' stories. This strategy identified well with some contemporary uses of narrative in art and design where the designer is delegated the role of providing narrative material to the audience, who then construct their own account of what is they think is going on. In practice however, the range of potential narrative options produced significant design problems. There was also a sense that there were predetermined narratives to be discovered which possibly narrowed the participation of the end user and made it questionable as a solution for engaging visitors.

Figure 3. Story Blocks

While these prototypes were used and discussed by the stakeholders the idea of a time travelling guide for women was also articulated as part of the meeting. This idea was discussed in relation to the interactive and material possibilities which were available from the prototypes. The outcome from this design meeting led then to a consideration of a third prototype which was based around the idea that end users or visitors could, through the assistance of authentic material objects, travel back in time to Cork in the 1700s.

5 The Time Travel Guide

The time travel guide, which is currently in development is designed around two key artefacts: a traditional guide book, and a travel case. The travel kit will be supported with historical factsheets which can be employed as part of the centre's education program. Aspects of this program may be delivered onsite through the education room (location C in Figure 1), or in the temporary exhibition space (location B in Figure 1), or through an outreach program to be developed with schools in the region.

The time travelling lady's guide to 1770s Cork

The travel guide focuses on a specific timeframe of the early 1770s. The Hibernian Chronicle newspaper was the main primary source for the historical information. The earliest archived editions start from this date. As with any city guide there are chapters that detail accommodation, places to eat, and things to do. Using information taken from newspapers and trade directories, the locations, people, and services featured in the book are factual. Unlike modern day guide books there are to be some details on etiquette and customs so the modern reader can blend in easily with her 1770s counterparts. The guide book will cater for those travelling on different budgets: large; medium and small.

The proposed featured chapters are:

1. Introduction - climate, currency, language spoken, and so on
2. Where to stay
3. How to decorate your home - the topic of taste was important in the 18th century
4. What to eat and where to buy it
5. What to wear
6. What to do/places to see
7. Hygiene
8. Health
9. Personal Safety
10. Tips for those who wish to work while on holiday

Figure 4. Prototype pages from the Travel Guide

The design of the book aims to reflect and give a flavour of eighteenth century style while remaining firmly a product of the 21st century. The eighteenth century will be suggested by the colour palette, visual style, and typography as indicated in Figure 4. The research will also consider the use of *inserts* to mimic *extra-illustrated books*. This was a form of scrapbooking which thrived in the eighteenth century where people would find a print, piece of artwork, letter, or other piece of memorabilia that was relevant to the text of a certain book, then inlay it in the original text and rebind the pages.

By placing the emphasis on visual communication rather than the traditional text-based method of conveying information this research aims to appeal to contemporary audience sophisticated in interpreting images (Mitchell, 1994, p.15). By using a variety of visual communication tools such typography, graphs/data visualisation, photography, illustration and visual puns this work aims to increase the reader engagement (Barnard, 2005, p.15, Klanten et al., 2011, pp.3-7). By harnessing the visual language of colour and typography to create distinct moods and settings, this visual design can portray and underline contrasts between the rich and poor of the city (Salen, 2000, p.79, Knight and Glaser, 2012). By featuring patterns and motifs taken from the decorative arts the project intends to produce a book which reflects beauty and femininity as it was understood in 18th-century Cork.

The narrative aspects of the book propose to work on two levels. Firstly, the book will feature snippets and newspaper clippings relating stories of people from the 1770s. Secondly, the guise of a travel guide will take familiar twenty-first century activities and transpose them back in time, thus encouraging the reader to picture themselves in the 1770s. By imagining how they might feel when faced with a certain situation or activity the reader can take facts from the page and continue the narration themselves, or in a shared context with others.

The time travelling lady's case – a museum in box

The contents of the book may be reconfigured into a series of factsheets and work modules which will accompany the case to be transported to schools for workshops, or used on-site in Nano Nagle Place. The suitcase is designed to contain eight individual boxes each relating to the chapters of the

book. There will be worksheets to accompany each box and suggested techniques for teachers on how to run a session with pupils. The eight boxes will cover: Home; Food; Appearance; Entertainment; Work; Hygiene; Health; and Law & order. Each of the boxes will contain three separate items and some examples are indicated in Figure 5. Table 1 give a list of objects currently under consideration. These number of items in the case make the case suitable for class group of 24 pupils and the case will take on the appeal of a museum in box. Each of items will be chosen to excite curiosity and playfulness; or their purpose might seem mysterious at first. Collaborative research is being carried out with other researchers in CIT whose field of research is focused on playfulness interfaces for museum settings.

Figure 5. Prototype material items and information cards to support the educational workshops using the time travel case.

Currently some opportunities for embedding technologies in the case and/or the individual items are being considered as a means of augmenting the interaction with the objects. However, the authenticity of the objects is key to the experience of them, so the introduction of technology will need to be carefully tested. At least one item in each box will need to be used, tried on, tasted, or physically interacted with in some way. The inventory of objects for box are still being selected and there will be some design thinking invested in whether the suitcase can be packed up with different boxes depending on the profile or age group of the end users on a given occasion. Some items might not seem of interest at first but the stories behind them will enhance their meaning.

Home	Food	Appearance	Entertainment
Lumps of sheep/beef fat Tea cup Wall paper	Small axe Recipe for badger fricassee Drisheen	Stays Patch box with patches Burnt material	Gaming chips Stones Direction for dancing the minuet
Work	**Hygiene**	**Health**	**Law & Order**
Knitting needles Daily timetable for servant Ribbons	Leaves, newspaper, strips of linen Patterns Bourdaloue	Remedy for a cough Bowl for bleeding Illustration of small pox	Note from Bridget to Rodger about poison Key to Gaol Handkerchief

Table 1. Considerations for contents to be included in the suitcase.

Initial discussions with the class teacher have resulted in the following scenario being proposed. The children will be divided into groups and each group given a box to examine. After an agreed time they will be invited to present their box of objects and their ideas as to what to the objects might be used for to the rest of the class. After each group has presented and demonstrated there will be a chance for the teacher to show the worksheets which will contain information on the objects and stories related to people who used them. This will be presented with sensitivity. It will important not to invalidate the children's suggestions when introducing the historical facts. An important part of the worksheets and follow-up discussion will be the presenting of the modern equivalent object. This will provide participants with a chance to reflect on both the differences and similarities of daily life between now and then. Table 2 shows some suggested comparisons which are currently being explored.

1770s	2016
Lumps of sheep/beef fat Tea cup Wall paper	Light switch Tea bags IKEA Catalogue
Small axe Recipe for badger fricassee Drisheen	Bag of sweets Menu from local restaurant Own brand of frozen burgers
Stays Patch box with patches Burnt material	Cotton vest Plastic surgery Child labour
Gaming chips Stones Direction for dancing the minuet	Wii Discussion about gangs Video for watch me (whip nae nae)
Knitting needles Daily timetable for servant Ribbons	Career leaflets Children could write their own timetable Woollen hat
Leaves, newspaper, strips of linen Patterns Bourdaloue	Roll of toilet paper Wellington boots Public toilet sign
Remedy for a cough Bowl for bleeding Illustration of small pox	Cough cough sweets Calpol Pink ribbon for cancer
Note from Bridget to Rodger about poison Key to Gaol Handkerchief	Divorce Discussion about punishment Discussion about honesty

Table 2. comparison of 1770 objects with modern equivalents.

The time travelling case is due to be piloted with fifth and sixth class pupils in a primary school in February 2017 and observations will be conducted around its use this context. The outcomes of this study will be invested in a redesign and further development of the case before moving on to the development phase. This final phase aims to produce a robust design of the suitcase, its components, and supporting material so that the project will be engaging and durable in the hands of future end-users.

6 Conclusion

The prototypes outlined here should be understood as early stage iterations in an overall design methodology which has narrative at its centre. These design prototypes are intended to hook the user into a cognitive process of recovering a story from material artefacts. In many cases these are designed to support learning, but are focussed on narrative design principles that support engagement rather than pedagogical principles for education. While the project has learning in its peripheral view the evaluation of learning, where necessary will remain within the control of suitably qualified professional educational officers or teachers, who are presently seen as secondary users of the project.

Design challenges that have yet to be addressed involve judgments about what role the materiality of the object plays in the provision of narrative content and consequently the users engagement with it. There is also questions about the practicalities of developing the suitcase of objects for use within the context of the educational programme. It is intended that the testing of the work using small focus groups, and later teacher-moderated classroom based studies, will reveal issues around the durability and practicality of the project.

7 References

ANDERSON, S. Forward. In: HONIBALL, N. & HADJILOUCA, D., eds. Narrative in Practice Design Symposium Catalogue, 21 May 2011 2011 Central Saint Martins College of Art & Design. University of the Arts London, 27.

BARNARD, M. 2005. *Graphic design as communication*, London ; New York, Routledge, Taylor & Francis Group.

BARTHES, R. [1966] 1977. Introduction to the Structural Analysis of Narratives. *Image, music, text*. London: Fontana.

BRUNER, J. S. 1986. *Actual minds, possible worlds*, Cambridge, Mass., Harvard University Press.

BRUNER, J. S. 1990. *Acts of meaning*, Cambridge, Mass., Harvard University Press.

COATES, N. 2012. *Narrative architecture*, Chichester, West Sussex ; Hoboken, N.J., Wiley.

DANT, T. 1999. *Material culture in the social world : values, activities, lifestyles*, Buckingham ; Philadelphia, Open University Press.

FLAHERTY, T. R. A. A. 2011. *Story Map* [Online]. Available: http://storymap.ie/ [Accessed 21/07/2013].

FLUDERNIK, M. 2003. Natural narratology and cognitive parameters. In: HERMAN, D. (ed.) *Narrative theory and the cognitive sciences*. Stanford, Calif.: CSLI Publications.

GERRITSEN, A. & RIELLO, G. 2015. *Writing material culture history*, London, Bloomsbury Academic, an imprint of Bloomsbury Publishing Plc.

HADJILOUCA, D., HONIBALL, N. & NAGASAWA, Y. Introduction. In: NINA HONIBALL, D. H., ed. NiP 2013 Creative Symposium, 21 May 2011 2013 Central Saint Martins College of Art & Design. Hato Press, 18.

HARVEY, K. 2009. *History and material culture : a student's guide to approaching alternative sources*, London ; New York, Routledge.

HERMAN, D. 2003. *Narrative theory and the cognitive sciences*, Stanford, Calif., CSLI Publications.

HERMAN, D. 2004. Towards a Transmedial Narratology. In: RYAN, M.-L. (ed.) *Narrative across media : the languages of storytelling*. Lincoln: University of Nebraska Press.

HERMAN, D. 2010. Narrative theory after the second cognitive revolution. In: ZUNSHINE, L. (ed.) *Introduction to cognitive cultural studies*. Baltimore: Johns Hopkins University Press.

HERMAN, D. 2012. Editor's Column: Transmedial Narratology and Transdisciplinarity. StoryWorlds: *A Journal of Narrative Studies*, 4, vii-xii.

JENKINS, H. 2006. *Convergence culture : where old and new media collide*, New York, New York University Press.

KLANTEN, R., EHMANN, S. & SHULTZE, F. 2011. *Visual Storytelling: Inspiring a New Visual Language*, Gestalten.

KNIGHT, C. & GLASER, J. 2012. When Typography Speaks Louder Than Words. *Smashing Magazine*.

LABOV, W. 1972. *Language in the inner city; studies in the Black English vernacular*, Philadelphia,, University of Pennsylvania Press.

MACLEOD, S., HOURSTON HANKS, L. & HALE, J. 2012. *Museum making : narratives, architectures, exhibitions*, Abingdon, Oxon England ; New York, NY, Routledge.

MITCHELL, W. J. T. 1994. *Picture theory : essays on verbal and visual representation*, Chicago, University of Chicago Press.

PAGE, R. E. 2010. *New perspectives on narrative and multimodality*, New York, Routledge.

PSARRA, S. 2009. *Architecture and narrative : the formation of space and cultural meaning*, Milton Park, Abingdon, Oxon ; New York, NY, Routledge.

RIESSMAN, C. K. 1993. *Narrative analysis*, Newbury Park, CA, Sage Publications.

RYAN, M.-L. 1991. *Possible worlds, artificial intelligence, and narrative theory*, Bloomington, Indiana University Press.

RYAN, M.-L. 2004. *Narrative across media : the languages of storytelling*, Lincoln, University of Nebraska Press.

SALEN, K. 2000. Surrogate Multiplicities: In Search of the Visual Voice-Over. In: SWANSON, G. (ed.) *Graphic design & reading : explorations of an uneasy relationship*. New York: Allworth Press.

SPEED, C. 2009. *Walking Through Time* [Online]. Edinburgh. Available: http://walkingthroughtime.eca.ac.uk/ [Accessed 21/07/2013].

SPEED, C. 2010. *Remember Me* [Online]. Manchester. Available: http://www.futureeverything.org/festival2010/rememberme [Accessed 21/07/2010].

Exploring new collaborative approaches in engagement and education towards the creation and enhancement of future STEM careers.

Dr Kieran Delaney [1], Dr Alex Vakaloudis [1], Paul Green [1], Trevor Hogan [1], Prof. Achilles Kameas [2], Dr. Ioannis D. Zaharakis [2],
[1] Cork Institute of Technology, Ireland, [2] Computer Technology Institute & Press, Greece
Email: kieran.delaney@cit.ie

Keywords: STEM

Introduction

European countries need as potential employees youngsters who are able to think creatively. They must apply new knowledge in an effective way, become continuously competitive in a highly demanding working environment through constant self-monitoring and thus be able to stand up for all the challenges that work based learning brings. The ability to switch efficiently between different disciplines (such as STEM [1]) depends on effectively processing the various forms of information based on clearly defined target outcomes, expanding a broad repertoire of ICT communication, problem-solving and decision-making skills, and using the collective knowledge represented in networks, based on working environments.

These are becoming urgent challenges in Europe and globally as a growing deficit in 21st century skills needs to be overcome. A recent report, the 'Future of Jobs' released by the Word Economic Forum [1] highlights the need for cross-industry and public-private collaboration; this envisages "partnerships between multiple businesses, educational institutions and accreditation providers can result in an overall increase in the quality of the talent pool, at lower costs and with greater social benefits". Reaching this goal is a significant challenge requiring new innovative solutions. This paper presents an approach to leverage novel new technologies, such as Ubiquitous Computing (UbiComp) [2], Mobile Computing (MobiCom) and the Internet of Things (IoT), in combination with the use of Communities of Practice (CoP) [2] to deliver these much needed solutions.

The objective of the research is to seek to expand the scale of student transdisciplinary projects by enabling the creation of a much broader range of more ambitious collaborative projects. These multi-disciplinary "projects of common purpose" would be meta-level initiatives that build upon the existing landscape, employing connective engagement models, new technology and tools that currently do not exist in order to achieve this. Students' participating to create, implement and disseminate transdisciplinary 'Challenge' projects would increase their empathic intelligence and problem-solving skills. They would also achieve a very strong understanding of how to collaborate, create and produce to real-world requirement and learn deeply how to operate effectively in dynamic environments.

[1] STEM: science, technology, engineering, and mathematics

[2] A community of practice (CoP) is a group of people who share a common craft or set of skills and share knowledge of the executive of these skills. The concept was proposed in 1991 by Jean Lave and Etienne Wenger. It should be noted that Communities of practice are not new phenomena; they have existed for as long as people have been sharing their learnt experiences through reports and storytelling.

Figure 1. High level overview of the elements present in a Community of Practice for STEM topics, which uses problem-based learning approaches.

Concept and Approach

In education, numerous programs exist to promote STEM activities; many take the form of project-based science and innovation competitions. This creates a landscape that talented STEM students will navigate without great difficulty; however, this is a fragmented picture for many students and a difficult one for teachers. The competitive nature of these projects acts to limit involvement and most operate annually to relatively short timelines. This effectively resets these programs every year. Importantly, an emphasis upon STEM skillsets means that there is often a significant deficit in creative enablers in most of these programs.

Many successful examples do exist of project-based collaborations built by multi-disciplinary student teams that transcend this. Children in coding clubs (e.g. CoderDojo) mentor each other using programs like Scratch and have subsequently self-organized to build teams to compete in robotics competitions. Undergraduates in engineering courses collaborate with business students to create new product prototypes; often these projects emerge from undergraduate work and evolve towards competitive programs, and may then be reinvented as postgraduate projects.

Our research envisages the creation of a continuum effect that connects these types of projects together through a suite of tools that enables students of different disciplines, ages, background and abilities to work together on transdisciplinary projects. Along these lines, this initiative is designing meta-level solutions to link school, community and third-level activities together and should foster a model that looks to strongly broaden impacts from the current cohort (e.g. elite students competing in STEM competitions) to the entire ‘education’ population. The project will also provide tools, particularly technology, to support this; however, the creation of this toolset must empower the stakeholder group and this will be considered carefully. New technology is often rejected, so approaches where participants can help co-create the toolkit must be prioritised.

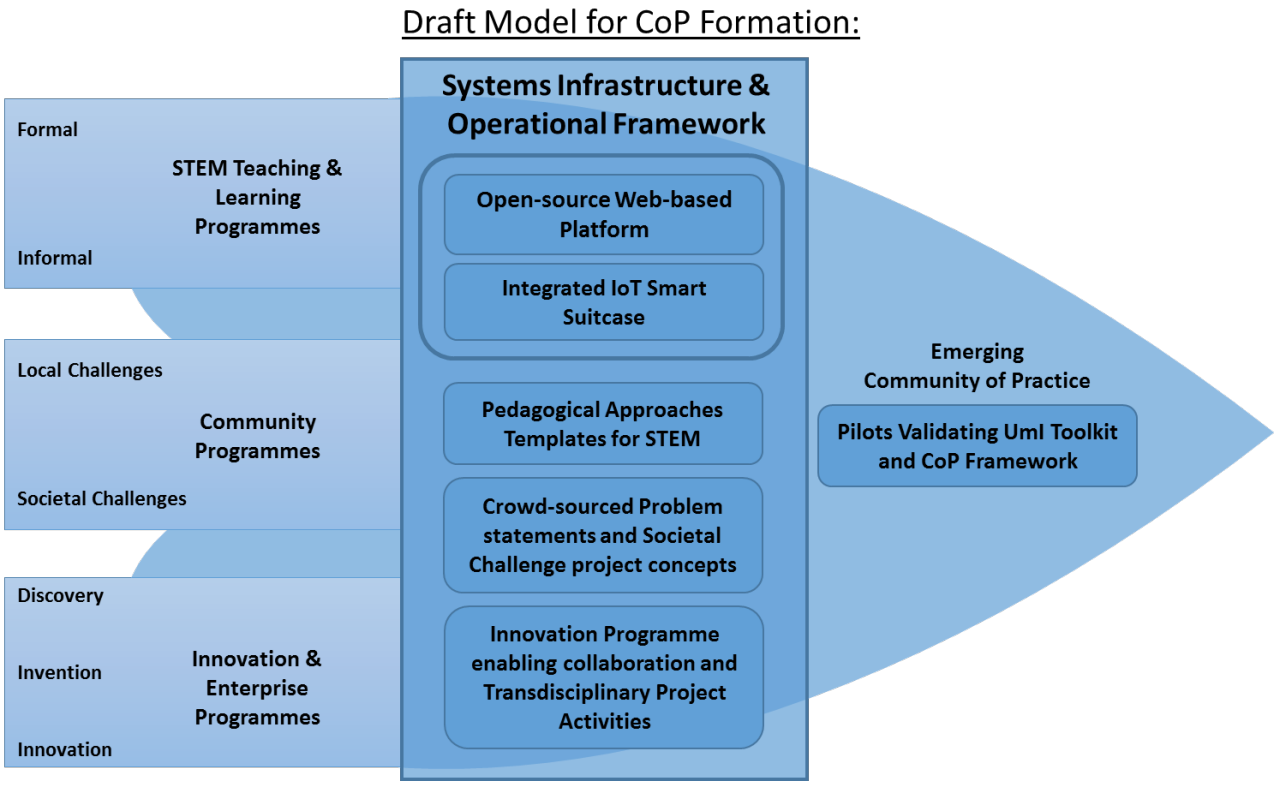


Figure 2. A Model for the Implementation of a Community of Practice: bringing Education, Industry and Communities together by using new enabling technologies and novel models of engagement.

The most effective way to do this will be to build these CoPs within active clusters where a body of knowledge already exists. These CoPs will need to foster cultures of innovation, investigate the drivers of success that can sustain key activities in the face of limited resources and high levels of turnover. [3, 4, 5]. A novel CoP framework will be presented to that explores new technologies and new models of engagement to create syntheses between STEM education, the arts and enterprise-driven and social innovation.

Conclusion

The central approach of this research will be multi-disciplinary project-based learning, with strong emphasis on increasing empathic intelligence and problem-solving skills. Student projects will reflect societal challenges, crowd-sourced from industry and communities in the innovation programmes and address challenging, complex issues; a mix of technology, arts and social science students will need to assemble to address them. Teachers will frame templates mapping the pedagogical requirements to project team profiles so that the integrity of assessments will be maintained. Students across many locations and disciplines will then self-organise to address the problems. A key innovation will be to foster scalable, collaborative internship projects operating across a range of host companies, mentored coherently by teachers from different disciplines and educational organisations.

Acknowledgements

This work is funded in part by the H2020 project UmI-Sci-Ed [710583]: Exploiting Ubiquitous Computing, Mobile Computing and the Internet of Things to promote Science Education

References

1. The 'Future of Jobs' report: http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf [accessed 12/08/2016]

2. Weiser, M. (1993). Ubiquitous Computing. Computer, 26(10), 71–72.

3. Lave, Jean; Wenger, Etienne (1991). Situated Learning: Legitimate Peripheral Participation. Cambridge: Cambridge University Press. ISBN 0-521-42374-0.; first published in 1990 as Institute for Research on Learning report 90-0013

4. Wenger, Etienne (1998). Communities of Practice: Learning, Meaning, and Identity. Cambridge: Cambridge University Press. ISBN 978-0-521-66363-2.

5. Wenger, Etienne; McDermott, Richard; Snyder, William M. (2002). Cultivating Communities of Practice (Hardcover). Harvard Business Press; 1 edition. ISBN 978-1-57851-330-7.

About UmI-Sci-Ed

UmI-Sci-Ed: Exploiting Ubiquitous Computing, Mobile Computing and the Internet of Things to promote Science Education [Project: 710583] Funded under H2020-SEAC-2015-1: Innovative ways to make science education and scientific careers attractive to young people

Many exciting new technologies are emerging, like Ubiquitous Computing (UbiComp), Mobile Computing (MobiCom) and the Internet of Things (IoT); in the following, we shall refer to them collectively as UMI. UMI technologies attempt to revolutionize everyday life by "allowing computers themselves to vanish into the background". UMI applications are presumed as a possible next generation computing environment in which each person is continually interacting with hundreds of nearby wirelessly interconnected devices: as a result, radical new uses of portable information technology emerge based on "the nonintrusive availability of computers throughout the physical environment, virtually, if not effectively, invisible to the user". These technologies are so modern and powerful that they can be both an educational means and end, thus fostering innovation and supporting promising scientific careers.

The broad aim of the project is to investigate the introduction of UMI technologies in education. By carefully exploiting state of the art technologies in order to design educational tools and activities, the project aims to offer novel educational services, implement innovative pedagogies and enhance students' and teachers' creativity, socialisation and scientific citizenship. We intend to put these technologies in practice, so as to enhance the level of science, technology, engineering and mathematics (STEM) education young girls and boys are receiving and at the same time make attractive the prospect of pursuing a career in domains pervaded by UMI. The orientation of UMI-Sci–Ed is entrepreneurial and multidisciplinary in an effort to raise young boys' and girls' motivation in science education and to increase their prospects in choosing a career in pervasive, mobile computing and IoT. We aim to develop an open fully integrated training environment that will offer 14-16 year old students and their teachers an open repository of educational material, educational scenarios, training material and activities, social tools to support communities of practice, entrepreneurship training, showcases, self-evaluation online tools, mentoring, and content and information management.

Workplace Well-Being: Mental and Psychological Health and Safety of Non-Unionised and Unionised Employees

Peter O'Mahony, Dr Angela Wright
Cork Institute of Technology, Ireland.
Dept. of Organisational and Professional Development. School of Business, Bishopstown, Cork.
Email: peteromahony@gmail.com, angela.wright@cit.ie

Keywords: Mental Health, Work Place, Guarding Minds at Work, Occupational Health, Psychosocial Risk, Workplace Well-being

Abstract: Imagine working in a highly productive environment in which you feel safe, respected and valued; the work is challenging; the demands of the job are reasonable; you have work-life balance; and your employer supports your involvement in your work and interpersonal growth and development. This is what is known as a mentally healthy workplace, (Canadian Centre for Occupational Health and Safety, 2012).

Twenty-three percent of workers in Europe are reported to have low levels of well-being (Ardito et al., 2012). Organisations, therefore, need to examine the operating environment so that they can begin to improve the psychological health of their employees (Mowbray, 2014). Failure to assess and resolve psychological health and safety risks in the workplace can create significant employee issues in an environment that requires "psychological competencies such as judgement, knowledge, interpersonal cooperation and emotional self-regulation. These psychological tools and skills flourish only in work environments that nurture and support their development and use, and minimize psychosocial factors in the work environment that can serve to undermine them" (Gilbert & Samra, 2010: para 2). While work can play a vital role in a person's quality of life as it provides them with a source of income and a platform for broader social advancement, it can also impact on a person's health as a result of risk factors present in the workplace which may cause injury, work-related illness or could potentially result in long-term health implications (Ardito et al., 2012). Consequently, mental and psychological health problems in the workplace can have a significant effect on both the employee and the organisation, in turn impacting productivity and competitiveness.

This research therefore seeks to explore the mental and psychological health and safety of employees within a multi-national pharmaceutical company based in the Republic of Ireland so as to address some of these pertinent issues. The participating company is listed in the top one hundred companies on the 2015 Fortune 500 list (Time Inc., 2015).

A quantitative case study was conducted in this study to gain a wider understanding of the mental and psychological health and safety of non-unionised verses unionised employees, specifically within this multi-national pharmaceutical company. One hundred and sixty-seven employees and seven senior managers were surveyed to explore their perspectives of psychological health and safety risks in the workplace across thirteen different psychosocial factors which are known to have either positive or negative effects on the psychological health and well-being of an employee.

To conduct this study, the research utilized online resources provided by Guarding Minds @ Work, an initiative which has been developed by the Canadian based Centre for Applied Research in Mental Health and Addiction (CARMHA). By using two online survey methods, this study tested a hypothesis which assumed that there would be more psychological health and safety risks among

unionised employees, while also investigating employee's experience of discrimination, bullying or harassment, and unfair treatment in the workplace. In order to test this theory, the following alternative hypothesis (H1) was tested:

'If a union exists within a workplace, then there will be a difference between the combined psychological health and safety concerns for non-unionised and unionised employees which will indicate that there is a greater risk to the psychological health and safety of unionised employees'.

Differences in the perspectives of all employees and senior managers were also compared with respect to the psychological health and safety of employees in the workplace. Statistical analyses in the form of t-tests were applied to the data obtained from ninety-seven non-unionised and sixty-three unionised employees to examine the level of psychological health and safety risks present within these groups.

The key findings from this study reveal that the psychological health and safety of unionised employees is more at risk in this workplace; in fact, unionised employees reported greater concerns across eleven out of the thirteen psychosocial factors. Furthermore, this research found that more unionised employees report previous experience of bullying or harassment in the workplace. It is also evident from this study that senior managers are significantly underestimating the psychological health and safety risks which are impacting employee's psychological health and safety in the workplace.

Considering that physically and psychologically healthy engaged employees maximise business efficiency and profitability, enabling organisations to achieve its goals and objectives, this study is pertinent and relevant to businesses who seek to protect the psychological health and safety of employees. As a result of dynamic workplaces, employees are now expected to be more flexible and resilient to changing organisational priorities as companies attempt to compete in challenging environments. Such challenges place a significant amount of pressure on employees which can inherently increase the risks to their physical and psychological health, safety, and well-being. Therefore, the main findings and recommendations for future practice and future research resulting from this study provide valuable insights not only for the participating multi-national company, but also for policy makers, employers, management, human resource departments, trade union representatives and professional or organisational development institutions.

References

Ardito, C., d'Errico, A., Leombruni, R., & Pacelli, Lia. (2012). Health and well-being at work: A report based on the fifth European Working Conditions, Dublin: European Foundation for the Improvement of Living and Working Conditions.

Canadian Centre for Occupational Health and Safety. (2012). Health & Safety Report, Vol. 10, No. 9, available at http://www.ccohs.ca/newsletters/hsreport/issues/2012/09/ezine.html, accessed January 9th at 6pm.

Gilbert, M, & Samra, J. (2010). "Psychological Safety and Health in the Workplace", Good Company, American Psychological Association for Organisational Excellence, Vol. 4, No. 5, May 12th.

Mowbray, D. (2014). "Psychologically healthy workplaces", The WellBeing & Performance Group, available at http://www.mas.org.uk/uploads/articles/Psychologically_healthy_workplaces_September_2014.pdf

Time Inc. (2015). Fortune 500, available at http://www.timeinc.com/experiences/the-fortune-500/, accessed January 6th, at 9pm.

Gender, Entrepreneurial Self-Efficacy, and Entrepreneurial Career Intentions: Implications of Female Role Models for Entrepreneurship Education

Ciara Lavelle O'Brien
Department of Management and Enterprise
CIT – Cork Institute of Technology, Ireland
E-mail: ciara.lavelleobrien@gmail.com

Keywords: entrepreneurship, self-efficacy, career intentions

Objectives

The objective of the research activity is to further the understanding of how entrepreneurship contributes to economic growth, competitiveness and social wellbeing.

Within the context of entrepreneurship education and entrepreneurial role models, this research explores the impact of gender. Guided by self-efficacy, entrepreneurship and entrepreneurship education theory three main research objectives are proposed. Firstly, to examine and assess how entrepreneurship education is delivered from primary school level education to third level education in an Irish context. Secondly, to explore the role and input of female role models in the entrepreneurship education system. Thirdly, to assess the effects of female role models on the entrepreneurial self-efficacy and career intentions of students.

Prior Work

This research draws on scholarly work from gender studies, entrepreneurship education research, role model theory and self-efficacy studies and includes writers such as Ajzen, I.; Bandura, A.; Henry, C.; Foss. L.; Kickul, J.; Wilson, F.; Marlino, D. and Barbosa, S. D.

In advanced market economies, women own 25% of all businesses and the number of women-owned businesses in Africa, Asia, Eastern Europe, and Latin America are increasing rapidly (Estes, 1999; Jalbert, 2000). This data reinforces the value of conducting further research into female entrepreneurship. (Women Owned Businesses, 2004).

The concept of female entrepreneurship is gaining traction in the academic world as well as the practitioner community. Indeed, Henry et al's (2012) review of extant gender and entrepreneurship literature's demonstrates a recent and significant proliferation of female entrepreneurship empirical research.

The Global Entrepreneurship Monitor (GEM), which provides useful international comparative information on entrepreneurship, reflects the difficulties for entrepreneurship which Ireland has experienced in recent years.

According to GEM 2013, Irish men are 1.9 times more likely than Irish women to be an early stage entrepreneur, with rates of early stage entrepreneurs at 12.1% for men and 6.4% for women. This has been steadily improving and although now level with the EU- 28 average of 1.9:1 and slightly higher than the OECD average of 1.7:1, it still shows untapped potential amongst female entrepreneurs in Ireland.

In 2014, the GEM survey confirmed again that entrepreneurial activity is mostly performed by men. More men than women are actively planning and starting new businesses in Ireland (2.1:1). Of the 20,400 individuals who started a business in Ireland in 2014, 14,400 are men (70.5%) and 6,000 are women (29.5%).

This situation is not unique to this country. Across G7 countries (Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States) a similar trend applies. In order to address this issue, the world's leading industrialised nations cited women's economic empowerment as a top global priority in a Joint leader's declaration presented on 8 June 2015 at the Group of Seven (G7) summit in Schloss Elmau, Germany and agreed the following common principles to support women's entrepreneurship:

- Make girls and women aware of the possibility of becoming entrepreneurs and actively encourage them to transform their ideas into business cases – starting at early stages, e.g. in schools, vocational training and universities by promoting tailored information.
- Counter gender stereotyping, and develop specific measures for girls to enrol in and complete education in STEM (Science, Technology, Engineering, and Mathematics) disciplines early on.
- Make successful female entrepreneurs more visible, e.g. as role models for new generations of women founders.
- Address the specific needs of female entrepreneurs, e.g. by providing them with tailored information, skills, mentoring and coaching and promoting networks for women entrepreneurs.
- Facilitate women entrepreneurs' access to finance, e.g. alternative sources of funding as well as the banking system, to technology and innovation and to domestic and international markets.

The declaration highlights the importance of catapulting women's entrepreneurship as a key driver of innovation, growth and jobs and provides further rationale for the need for research in this area.

Prior research has been conducted on entrepreneurial self-efficacy, gender roles in entrepreneurship as well as a large amount of research attention on entrepreneurship education. There does however seem to be a gap in combining these theories together to recognise the effect that entrepreneurship education has on female students in particular as well as examining the effect of female role models on the entrepreneurial mind-set and career intentions of students.

With regard to the methodology, the use of George Kellys (1955) repertory grid technique (RGT) in entrepreneurship research has been minimal thus far (Klapper 2015). This addresses another gap in the research.

The outcomes of this research will provide insights into how entrepreneurship is delivered in the modern classroom as well as the impact of female role models on the entrepreneurial mind-set and career intentions of students.

Approach

Extensive work has been conducted in an attempt to examine entrepreneurship education but has been inconclusive (Cox et al., 2002). This may be linked to methodological issues, specifically the outcome measures used in many studies, such as student satisfaction and performance, as these are insufficient indicators (Cox et al., 2002). For the purpose of this study a qualitative approach will be adopted as it has the potential to increase the validity, depth, richness and creativity of entrepreneurship education research. Qualitative research encompasses a variety of methods that can be applied in a flexible manner, to enable respondents to reflect upon and express their views. Qualitative data typically captures what OECD (2009) refers to as hard (actions taken such as start-up of a new venture) and soft (attitudes, intentions, cognitions etc., such as self-efficacy, intention to become an entrepreneur).

This research will explore the use of George Kellys (1955) repertory grid technique (RGT), the methodological tool of Personal Construct Theory (PCT). One of the main aims of PCT is to understand people's unique views of the world by exploring their thoughts, feelings and beliefs (Cooper, 1998). Data will be collected mainly by methods such as semi-structured interviews, participant observation and repertory grids, which are a methodological tool of George Kelly's Personal Construct Theory (1959).

Implications

This research will have practical as well as policy implications. From a policy perspective, the research will inform entrepreneurship, education and gender policy development. From a practical perspective, this research will provide empirical evidence of the role and impact of female role models across the education system.

Value

Due to the research deficit, it is anticipated that the ongoing study will make a substantial contribution to both academic knowledge and practice by expanding the entrepreneurship literature. The study will also highlight essential aspects of the discipline which warrant further investigation in the future. From a research perspective this research will also bring about a better understanding of how role models are constructed in the minds of entrepreneurship students.

References

Cox, L., Mueller, S., & Moss, S. (2002). The impact of entrepreneurship education on entrepreneurial self-efficacy. International Journal of Entrepreneurship Education, 1, 2.

Global Entrepreneurship Monitor Report (2013), Paula Fitzsimmons, Colm O'Gorman 2011, [Online] Available at: https://www.enterprise-ireland.com/en/Publications/Reports-Published-Strategies/GEM-Reports/GEM-Report-2013.pdf

Global Entrepreneurship Monitor Report (2014), Slavica Singer, Jose Ernesto Amoros, Daniel Moska Arreola 2014, [Online] Available at: http://www.babson.edu/Academics/centers/blank-center/global-research/gem/Documents/GEM%202014%20Global%20Report.pdf

Playing in the museum: shaping a definition of digitally mediated playful experiences

Denise Heffernan [1], Dr Kieran Delaney [1], Paul Green [2]
[1] Nimbus Centre, [2] Department of Media Communications
CIT – Cork Institute of Technology, Ireland
Email: denise.heffernan@mycit.ie

Keywords: museum, play, embodied interaction

Introduction

Museums have always been early adopters of new technology and in recent years, museums have increasingly embraced gameful and playful design (Rugged Rovers; Designing the Pen, 2014). However, museums traditionally have a preference for a didactic approach and emphasize learning over personal experience. Digital technologies have been embraced typically with the aims of democratising knowledge, contextualising information, and ultimately boosting visitor numbers. Research on promoting play within the museum has been focused on educational goals (Taylor et al., 2015). Little consideration has been given to designing for the intrinsic value of play; play itself.

Play is a form of understanding who we are, what surrounds us and how to engage with others (Sicart, 2013); the social context for playing is particularly important in creating memorable experiences. Play is not outcome-driven but focused on the process. It is inherently driven by curiosity. It encourages exploration and experimentation, results in discoveries, is often spontaneous, and is not concerned with goals. An environment can encourage and promote play just as it can hinder play. Playfulness is distinctly different to game-based activities. The adoption of gameful methods don't encourage open-ended play, but a specific goal-based learning interaction.

Emerging technologies that underpin the Internet of Things provide an opportunity to explore a space without the limits associated with typical interactive devices found in the museums setting. We speculate that the potential for a connected network of embedded digital interventions could allow the visitor to playfully investigate, explore and re-interpret their museum experience. In this paper we explore the design of playful, digital meditated experiences. We describe the design process that is being undertaken in this research. We discuss our play matrix as a tool that helps define playful activities. We briefly review digitally augmented museum experience, and finally, we reflect on the importance of this research and future directions.

Related Work

Research within the field of interaction design (IxD) and human computer interaction (HCI) has produced a variety of technological solutions for the museum context which are mobile, interactive, and reusable. Projects such as the MeSch Project (Material EncounterS with digital Cultural Heritage) aim to provide a series of tools that will allow cultural heritage professionals design and implement interactive experiences without the need for "expert" developers; supporting the creation of an open [playful] community of cultural heritage professionals driving and sharing a new generation of interactive physical/digital smart exhibits. (Petrelli & Lechner, 2014).

While the education goals of museums are still paramount, there has been a shift from the didactic model to a more adaptive, socially inclusive dialogical approach. This coupled with the emerging DIY approach afforded through projects like MeSch, moves the museum from teacher to facilitator; "a place where visitors can create, share, and connect with each other around content" (Simon, 2010). A place where playful exploration becomes central to a more engaging and memorable experience.

Research approach

The research has adopted a mixed methods approach. The focus of this initial investigative stage is to define a hierarchy of playful criteria and sensory modalities most appropriate to play. This involved a survey of literature as well as a review of current playful and technological practices employed by museums.

We began by defining initial play criteria through a literature review of play theories. This formed the basis of our play matrix, a tool to identify observable playful behaviour. The play matrix helped to observe and measure play activities. We tested the matrix in museums and other public spaces, including playgrounds where overtly playful behaviour is found. Observations using our play matrix, noted that play activities often in their appropriation of the space was increasingly tactile and social. Sticks become swords. Climbing frames become mountains. Acquaintances became trusted allies. The evaluation of the matrix allowed dominant and new criteria to emerge and began to reveal a hierarchy of sensory modalities which could be mapped back to our playful criteria. Emerging sensory modalities will be used as a method of interpreting how IoT technologies should be built to genuinely enhance opportunities for playfulness.

Self-directed	Imaginative	Experimentation	Social	Collaborative	Discovery	Challenge	Body Language
Minimal direction from museum/staff/adults	Separate from real world restrictions	Using the space/object for something other than intended	Engage with others in own group/family unit	Asking another to join activity	Questioning	Requires application of self	Open - Welcoming, friendly, Inclusive
Freedom of choice (leaving the prescribed route etc)	Role playing	Observation, watching and learning	Engage with strangers	Combine with another to finish/start an activity	Looking beyond what is presented	Does not cause anxiety	Active
Choose to engage, not forced	Prescribing roles to another	Test boundaries & limitations	Communication outside of expected interactions	Including others in an activity	Searching	There is some resolution to the challenge	Animated
Stopping/Starting at own discretion	Engrossing	Freedom from externally imposed rules (e.g quiet in the museum)	Laughing	Joining others in their movement through the space	Finding, prescribing new meaning	Problem solving	Inquisitive, picking up objects, getting closer to an object
Non instruction led activities	Fantasy/ make believe elements		Cross-generational engagement	Cooperation with others	Attention to the activity rather than outcome		
	Active engagement				Curiosity		

Figure 1. Example of initial play matrix

A review of current practices in museums was also undertaken through a review of literature and in-the-field observations. We found that technological solutions often cut visitors off from the social interactions that define and shape the museum experience. The museum visit is shaped and enhance through our encounters with those whom we visit with and encounter within the museum (Falk & Dierking, 2012). However, there are many examples of technology that hinder social, engaging interactions: the popular use of audio guides can isolate the visitor, actively discouraging talking and engaging with one another; a kiosk directs attention towards itself and away from artefacts on display, and interactive interfaces and touchscreens that are led by a single user. Designing for playful engagement could break down the barriers often associated with current technological solutions and, importantly, create a more memorable social experience.

Discussion

The possibilities offered by emerging IoT technologies and playful design can help support the museum to offer new experiences, even to an audience already familiar with its content. There is opportunity to take advantage of visitors' physical experience with cultural heritage and to integrate technology into it instead of creating a parallel and detached digital experience. It also provides the opportunity to create a social and collaborative experience with those we visit with and encounter in the museum. Research not specific to exhibition design has shown how embedded technologies can be used to create, meaningful, collaborative, social experiences (Speed, 2010). Technology can be used to encourage playful, exploratory modes of engagement, and to provide an experience-centric not learning-focused museum visit.

By focusing on designing for the social, exploratory nature of play could support a more meaningful, social experience for some within museums when compared to other activities such as tours or presentation material that are prescribed by the museum curator. We have begun to outline design sensitivities, the next steps are to evaluate and validate these design sensitives and further explore what opportunities and limitations do IoT technologies have in the implementation and support of playful spaces in complex public environments such as the museum.

Future Direction

By focusing on designing for the social, exploratory nature of play could support a more meaningful, social experience for some within museums when compared to other activities such as tours or presentation material that are prescribed by the museum curator. We have begun to outline design sensitivities, the next steps are to evaluate and validate these design sensitives and further explore what opportunities and limitations do IoT technologies have in the implementation and support of playful spaces in complex public environments such as the museum.

References

Cooper Hewitt Smithsonian Design Museum. (2014). Designing the Pen. [online] Available at http://www.cooperhewitt.org/new-experience/designing-pen/ [Accessed 28 Nov. 2015]

Falk, J. & Dierking, L. (2012). Museum Experience Revisited. Walnut Creek, CA, USA: Left Coast Press.

Lechner, M. & Petrelli, D. (2014) The Mesch project: reusing existing digital resources in the creation of novel forms of visitor's experiences in CIDOC - the International Committee for Documentation, Annual Conference - Access and Understanding: Networking in the Digital Era. Dresden

Preloaded Games. (n.d.). Rugged Rovers. [online] Available at http://preloaded.com/games/rugged-rovers/ [Accessed 28 Nov, 2015]

Sicart, M. (2014). Play Matters. MIT Press.

Simon, N. (2010). The Participatory Museum. Santa Cruz. Left Coast Press, Inc.

Speed, C. (2010). An Internet of Old Thing. Digital Creativity, 21(4), 235–246.

Taylor, R., Bowers, J., Nissen, B., Wood, G., Chaudhry, Q., Wright, P. (2015). Making Magic: Designing for Open Interactions in Museum Settings in Creativity and Cognition Conference, Glasgow. pp. 313-322

Meeting in the Middle:
A collaborative and multidisciplinary approach to developing an empowerment course for women survivors of gender-based violence.

S. Davis, B. Kenny
Hincks Centre for Entrepreneurship Excellence,
Department of Management and Enterprise
CIT – Cork Institute of Technology, Ireland
e-mail: sarah.davis@cit.ie

Keywords: Collaborative learning, Gender-based violence, Women's empowerment

Introduction

The purpose of this paper is to explore the challenges faced in the creation and piloting of a training programme for women survivors of gender-based violence in Ireland. The training programme included employability and entrepreneurship skills and combined this with elements of personal-development to improve self-confidence and self-efficacy. Empowerment of women who had experienced gender-based violence was the goal of the Daphne III project called NEW START which ran from 2014-2016 in seven EU countries.

Gender-based violence (GBV), also commonly known as domestic violence (McWilliams and McKiernan 1993), is a complex and wide spread phenomenon. In Ireland, over 12,500 people annually, 9,448 women with 3,068 children received support and/or accommodation from a domestic violence service (SAFEIreland 2016). Women who have experienced GBV tend to have low self-esteem and low self-efficacy. This paper draws on the literature to develop an understanding of the collaborative process in curriculum development and seeks to reconcile this with women's empowerment literature and an understanding of GBV survival and recovery. The Wheel of Power and Control (Goodman and Fallon 1994) a commonly-used model to explain GBV to those seeking assistance and support is reviewed. An understanding of power and empowerment is then sought (Charmes and Wieringa 2003) and finally, the educational literature in the areas of andragogy, heutagogy (Blaschke 2012) and transformational learning (Mezirow 1991) is explored.

Design/Methodology/Approach

This paper outlines the collaborative development (Voogt, Westbroek et al. 2011) and implementation of a ten week pilot programme that was delivered in Ireland in 2016 to eight women who had previously experienced domestic violence. The EU project requirement was to combine elements of life-coaching and elements of mentoring or employability training with personal development training and, to produce and test training materials suitable for use with women survivors. A qualitative emergent enquiry approach (Keegan 2009) was taken (by the collaborating partners in Cork Institute of Technology and YANA (You Are Not Alone) Domestic Violence Project) to the multidisciplinary programme development and implementation. Qualitative data in the form of notes from training the trainers and notes from the pilot sessions debriefing were analysed. Thematic analysis was used to identify the main themes from this qualitative data.

Quantitative data, in the form of three standard instruments for self-esteem (Rosenberg 1965), self-efficacy (Schwarzer and Jerusalem 2010) and entrepreneurial self-efficacy, were gathered in a pre- and post-test with the women (N=7) who completed the pilot programme. The women were all between 35 and 65 years of age and all were current clients with YANA, a north Cork domestic violence project with outreach services. All had attended the 'Pattern Changing' programme (Goodman and Fallon 1994) or had received personal development training prior to undertaking the NEW START programme. The aim of this quantitative test data was to establish whether the training had been effective in empowering the women by improving their self-esteem, self-efficacy and entrepreneurial self-efficacy.

Findings

The findings of this research show that there is benefit in extending the training available to survivors of GBV by bringing in outsider trainers and by covering mainstream employability and entrepreneurship training materials. Also, for this cohort of women during this pilot, the benefit was mostly in the training process rather than in the mastering of the material covered. Similar training was provided by the other European partners. (*– awaiting final report to insert overall result summary*) In general, demand for entrepreneurship in the context of starting up a business was limited, but entrepreneurial skills were relevant.

Research Implications/Limitations

The findings are based on a single implementation of the programme. The researchers were actively involved in all stages of the project. However, the results obtained were in line with those found by the six other European partners contributing to the NEW START program. However, due to the limited number (N=7) who received training in Ireland and only one group implementation, no generalisable conclusions can be drawn. The programme developed should be tested with further groups to establish reliability of this means of empowerment.

Originality/Value

The added value of the Irish NEW START course contribution is that it is a programme that directs itself to the subject of 'moving on' from being defined by the experience of GBV. This programme was in effect a transitional bridge between the safe and known environment of an outreach provider and the greater world. It provided a space where the women could still check-in while also checking out other wider options.

References

Blaschke, L. M. (2012). "Heutagogy and lifelong learning: A review of heutagogical practice and self-determined learning." The International Review of Research in Open and Distributed Learning **13**(1): 56-71.

Charmes, J. and S. Wieringa (2003). "Measuring women's empowerment: an assessment of the gender-related development index and the gender empowerment measure." Journal of Human Development **4**(3): 419-435.

Goodman, M. S. and B. C. Fallon (1994). Pattern changing for abused women: An educational program, Sage Publications.

Keegan, S. (2009). "Emergent inquiry." Qualitative Market Research: An International Journal **12**(2): 234-248.

McWilliams, M. and J. McKiernan (1993). Bringing it out in the open: domestic violence in Northern Ireland, HM Stationery Office.

Mezirow, J. (1991). Transformative dimensions of adult learning, ERIC.

Rosenberg, M. (1965). "Rosenberg self-esteem scale (RSE)." Acceptance and commitment therapy. Measures package **61**.

SAFEIreland (2016). The State We Are In In 2016: Towards a Safe Ireland for Women and Children. Athlone, Co. Westmeath, SAFE Ireland.
Schwarzer, R. and M. Jerusalem (2010). "The general self-efficacy scale (GSE)." Anxiety, Stress, and Coping **12**: 329-345.

Voogt, J., H. Westbroek, A. Handelzalts, A. Walraven, S. McKenney, J. Pieters and B. de Vries (2011). "Teacher learning in collaborative curriculum design." Teaching and Teacher Education **27**(8): 1235-1244.

A Serious Game to Dissuade Adolescents from Taking Performance Enhancing Drugs

Larkin Cunningham
Department of Computing
CIT – Cork Institute of Technology, Ireland
e-mail: larkin.cunningham@cit.ie

Keywords: Serious Games, Education, Performance Enhancing Drugs

Can a serious game that invites players to explore the morality of performance enhancing drugs (PEDs) from several perspectives (athlete, federation, wider society) be effective in changing the attitudes of athletes and non-athletes to PEDs?

Performance enhancing drugs (PEDs) are a major issue not just in professional sports, but also in amateur and recreational sports, including the use of androgenic anabolic steroids (AAS) in gym culture (Hartgens and Kuipers 2004). Studies have shown that between 7% and 12% of young athletes would choose to dope using an illegal substance if it guaranteed improved performance, could not be detected and did not adversely affect lifespan (Bloodworth *et al*. 2012; Connor *et al*. 2013).

The abuse of non-prescribed PEDs, such as AAS, is a public health concern that requires the implementation of education programmes aimed at adolescents (Tahtamouni *et al*. 2008); research has shown that such programmes can be effective in reducing the intentions of adolescent athletes to use supplements (DesJardins 2002).

Michael and Chen (2005) offer the qualified definition of serious games as "game[s] in which education (in its various forms) is the primary goal, rather than entertainment", though they do concede that games intended primarily for entertainment purposes can contain serious game elements. Serious games have seen rapid growth for more than a decade in both academia and industry, in areas such as education and cultural heritage, well-being and health care, and they have huge potential in improving achievements in these areas and others (Laamarti *et al*. 2014).

One existing serious game, Play True Challenge, was developed for the 2010 Youth Olympic games (World Anti-Doping Agency 2016). It is a small-scale cartoon-like 2D adventure game featuring a sport similar to steeplechase at its core. Players learn about banned substances like erythropoietin (EPO), steroids and stimulants; they can take these in the game and discover the consequences later, such as being dropped from the athletics team.

A new serious game is under development, aimed primarily at adolescents, but also older athletes and non-athletes, to inform them about the consequences of taking PEDs. It will be more extensive than WADA's Play True Challenge and use updated technology to explore a wider range of moral grey areas of PED usage. Its primary contribution will be an examination of the efficacy of a serious game in bringing about attitudinal change with respect to athlete doping and the tolerance of wider society to the use of PEDs.

Using Unreal Engine, the engine used in commercial successes like Gears of War and Tom Clancy's Rainbow Six Vegas, the game is being developed using an immersive 3D environment in the first-

person perspective. Learning outcomes of the game are being informed by literature of the type presented in this short paper, as well as other educational materials on the positive and negative effects of PEDs. A number of popular game mechanics are being employed, such as object interaction and collection, goal completion, challenges, puzzles, achievements and dynamic interactive dialogue.

Ultimately, the game is intended to be a persuasive one. Ian Bogost, in his book Persuasive Games (2007), defines procedural rhetoric as “the art of persuasion through rule-based representations and interactions, rather than the spoken word, writing, images, or moving pictures”. His interest lies in how videogames argue about the way systems work in the real world outside the videogame. A persuasive game on the subject of PEDs would, therefore, through the medium of the 3D first-person game under development, represent the reality of doping in sport, how it can be nefarious and be protected by corruption, what the daily reality of doping can be like for an athlete, the devastating long-term effects it can have on personal health and reputation.

The main protagonist in the game (the primary player character) is an investigative journalist piecing together the story of doping in sport for an exposé article. He interacts with a number of adolescents with hopes of future stardom, existing athletes and retired athletes (see Fig. 1 for an example of this interaction through dynamic dialogue). The game will feature some flashbacks from the perspective of the athletes, where you as the player will make choices, such as whether to take a banned substance or not, and see how this affects the athlete’s story in the present. This will have a dynamic effect on the final exposé article.

The game will be modular in nature, allowing for it to be customized for different sports if required. Characters can be swapped for others, dialogue can be edited separately from the game engine code, it can be localized for different languages, and so on.

Development of the game is underpinned by the established Game Object Model (GOM) version II theoretical framework (Amory 2007), which uses the object-oriented programming metaphor. The model describes a number of interfaces to game elements that can be abstract (pedagogical and theoretical constructs, such as critical thinking, emotiveness, goal formation and goal completion) or concrete (design elements, such as story, plot, backstory, graphics and sound). These are then mapped to the intended learning outcomes and the game mechanics.

Aside from outlining the background of this project, this short paper is also a call for collaboration. The development of a persuasive videogame set in an immersive 3D environment is an ambitious and multifaceted one. Researchers from a variety of backgrounds, such as pedagogy, psychology, sport science, digital arts, game studies and other areas could contribute.

In conclusion, there is an obvious problem with PEDs, made all the more topical by recent scandals, such as in athletics (Russia banned from athletics in 2016 Olympics), tennis (Maria Sharapova banned for two years) and mixed martial arts (Brock Lesnar testing positive after earning $2.5 million for a fight at UFC 200). Education on the subject of PEDs and other supplements has been shown to be effective. Serious games are a growing force in education and have the potential to greatly improve educational outcomes. A persuasive serious game can show how systems in the real world work, preparing the adolescents who play the game for the choices they will face and be aware of the consequences of their choices.

Fig. 1 – A screenshot showing dialogue from an early prototype

References

Amory, A. (2007) 'Game Object Model Version II: A Theoretical Framework for Educational Game Development', *Educational Technology Research and Development*, 55(1), 51–77.

Bloodworth, A.J., Petróczi, A., Bailey, R., Pearce, G., McNamee, M.J. (2012) 'Doping and supplementation: the attitudes of talented young athletes', *Scandinavian Journal of Medicine & Science in Sports*, 22(2), 293–301.

Bogost, I. (2007) *Persuasive Games: The Expressive Power of Videogames*, MIT Press: Cambridge, MA, USA.
Connor, J., Woolf, J., Mazanov, J. (2013) 'Would they dope? Revisiting the Goldman dilemma', *British Journal of Sports Medicine*, 47(11), 697–700.

DesJardins, M. (2002) 'Supplement use in the adolescent athlete', *Current Sports Medicine Reports*, 1(6), 369–373.

Hartgens, F., Kuipers, H. (2004) 'Effects of Androgenic-Anabolic Steroids in Athletes', *Sports Medicine*, 34(8), 513–554.

Laamarti, F., Eid, M., Saddik, A.E. (2014) 'An Overview of Serious Games', *International Journal of Computer Games Technology*, 2014.

Michael, D.R., Chen, S.L. (2005) *Serious Games: Games That Educate, Train, and Inform*, Course Technology / Cengage Learning: Boston, MA, USA.

Tahtamouni, L.H., Mustafa, N.H., Alfaouri, A.A., Hassan, I.M., Abdalla, M.Y., Yasin, S.R. (2008) 'Prevalence and risk factors for anabolic-androgenic steroid abuse among Jordanian collegiate students and athletes', *The European Journal of Public Health*, 18(6), 661–665.

World Anti-Doping Agency (2016) Play True Generation [online], available: https://www.wada-ama.org/en/play-true-generation [accessed 25 Jul 2016].

A User Engagement framework for Universal Inclusion

A User Engagement framework for Universal Inclusion
Alex Vakaloudis, Mark Magennis, Martina Hayes, Jane O'Flynn, Kieran Delaney
Nimbus Centre. CIT – Cork Institute of Technology, Ireland
e-mail: mark.magennis@mail.com, martina.hayes@mycit.ie, {janeoflynn,alex.vakaloudis, kieran.delaney}@cit.ie

Keywords: User Engagement, social inclusion, cross-sectoral project management

Introduction

Numerous innovative products, albeit pioneering, have not been successful because they have not being adopted by their end-users as they disregard their functional or interfacing needs. Likewise, the wishes of end-users are not easily or effectively communicated to researchers and industry innovators who can make use of technological advances so as develop prototypes and eventual products to correspond to these needs. The solution is to create a sustained process that engages at an early stage the cross-sectoral characteristics of a project and involve all stakeholders in fruitful and effective collaborations. In this paper we summarise the development of User Engagement Frameworks design to achieve this in a sustained manner.

Related work

There is a huge increase in the involvement of communities of people in cross-sectoral projects. There are significant drivers in the academic world (national civic engagement targets) [1][6], the corporate world (emphasis on CSR benefits) [4][5] and the public sphere (push for more effective forms of public engagement) [2][3].

Concepts and Approach

The aim of this effort is to provide the tools so as to support the management of cross-sectoral collaborative projects where at least one partner is an academic institution and the other partner(s) may be from the academic, corporate or public sectors. The end-users need to be involved in all stages of the project lifecycle from requirements elicitation through to specification of user-experience to proof of concept and prototype trialling. Likewise industrial stakeholders encompass a pivotal role to realistic product development and possible commercialisation.

Apart from offering project management, the main objectives are to improve project outcomes by reducing the risk of one or both partners not achieving their goals, to save time and cost by allowing partners to store and re-use successful project processes and supporting materials and to enable organisations to develop their project management capabilities to benefit similar future projects.

Methodology

We used interviews and focus groups to ascertain each stakeholder's requirements for how a resulting collaborative platform should work and what functionality it should offer. The unique challenges in cross-sectoral engagements are:

- Finding the right opportunities and partners based on their skillsets, availability and compatibility
- Understanding each other's aims and expectations and achieving a shared vision of the goals of the project
- Understanding what each partner can contribute and under which conditions
- Organising activities, working together effectively and managing interactions

A platform to meet these challenges consists of the following functional modules

- User Management Module to manage the profiling attributes of participating organisations and participants and assign appropriate access rights for each project
- Dialogue Module to implement messaging among team members' preferred channels (e.g. email, SMS, WhatsApp, Facebook) in order to schedule and realise (e.g. interaction activities such as discussions, meetings and interviews and the creation of demos/prototypes of any level of maturity. Furthermore to give a mechanism for everyone outside academia to express needs and research ideas.
- Project Management Module to create a project agenda, allocate tasks including any documents (e.g. the informed consent of the participants) and organise activities with relevant information such as aims, date, time and location, required participant demographics, abilities/disabilities, digital skill level, use of assistive technologies, level of compensation offered, etc.
- Dissemination Module to provide metrics on the user engagement.

Case Studies

Two case studies implementing this work will be described:

ASPIRE (Application to Support Participant Involvement in Research) is a programme for universal inclusion targeting projects that involve people with disabilities. The intention of ASPIRE is to support their inclusion in research, co-design, social campaigns and other activities, regardless of their physical, social or economic abilities. Apart from identifying the functional modules for platform to fulfil these needs, ASPIRE highlighted the cyber-physical characteristics of such interaction and identified the importance of multiple and novel interaction methodologies such as smart/wearable devices and smart surfaces.

Melu is a web-based platform for cross-sectoral projects that enables detailed and flexible member profiling with skills, capabilities, conditions and requirements in order to achieve optimum matching for collaborations and activities. Apart from managing projects, Melu aims to build a knowledge base of procedures/materials to improve the formation of successful collaborations.

Conclusions and future work

This is a system that takes the pain out of cross-sectoral project management, improves project outcomes, reduces the risk of one or both partners not achieving their goals and saves time by allowing partners to save and re-use successful project processes and supporting materials. There is scope for additional future work. This refers to the broad adoption of IoT technologies by understanding and overcoming practical and cultural barriers and the detailed customisation of the interface depending on the micro-profiling of participants,

Acknowledgements

ASPIRE was developed as part of funding by the Irish National Council under the New Foundations programme in collaboration with NCBI.

References

1. ViPi (Virtual Portal for Interaction and ICT Training for People with Disabilities) KA3 LLL project Lifelong Learning program, subprogramme KA3 ICT (511792-LLP-1-2010-1-GR-KA3-KA3NW). https://www.vipi-project.eu.

2. e-LIFE (e-Learning Virtual Center for People with Physical Disabilities) European Union, Lifelong Learning Programme 2007-2013, Leonardo Da Vinci Sub-Programme, Transfer of Innovation Code: 2011-1-RO1-LEO05-15322. http://www.elife-eu.net.

3. University of Cambridge. Inclusive Design Toolkit. http://www.inclusivedesigntoolkit.com

4. IDC, Worldwide Quarterly Mobile Phone Tracker http://www.idc.com/tracker/showproductinfo.jsp?prod_id=37

5. http://www.doro.ie/

6. J.Ye, G. Stevenson, and S. Dobson (2014). KCAR: A knowledge-driven approach for concurrent activity recognition. Pervasive and Mobile Computing, 201

Building capacity when assessing non-formal and informal learning through collaborative networks for Recognition of Prior Learning

P. O'Leary, L. Hearne, A. Ledwith

Cork Institute of Technology, University of Limerick, Ireland

phil.oleary@cit.ie, l.hearne@ul.ie, a.ledwith@ul.ie

Keywords

Recognition of Prior Learning, Non-formal and informal learning, values, beliefs, assumptions

Abstract

The current European policy climate is intended to create conditions where Recognition of Prior Learning (RPL) can flourish within the formal learning system. However, RPL is a complex process to deliver in practice given the diverse nature of non-formal and informal learning and its assessment.

It is recognised that cultural acceptance of RPL across an institution is a critical factor cultivating conditions for pedagogic agency for RPL (Cooper and Harris, 2013). This is true for each of the actors and particularly for academic staff. This research focuses on the Irish higher educational institutional system in particular and takes a novel approach to supporting RPL provision by exploring the values, beliefs and assumptions of assessors and candidates in order to better understand the perspectives of both actors (Friesen, 2011). This research has taken a constructivist grounded theory approach to explore the perspectives at play within RPL (Bryman, 2012, Charmaz, 2006). The setting for the research is within one third level institution in Ireland. Semi-structured interviews were held in 2015 with 31 academic assessors. In 2016, 27 past candidates of RPL were interviewed. Both interview data sources were asked about the values, beliefs and assumptions inherent in their understanding of the RPL process.

Aspects of Bernstein's theories (2000) have provided the theoretical framework to support preliminary analysis of data, including the idea of the classification of knowledge, the totally pedagogised society, and the field of recontextualisation where knowledge is repositioned to facilitate entry to the formal learning system. Results so far in this study reveal that the academic assessor looks first to maintain the institutional standards, however they are willing, where merited, to step towards supporting the candidate's claim, revealed as the act of 'balancing.' Results also show that candidates possess values, beliefs, and assumptions that strongly defend their reasons for applying, and the validity behind the basis of their case. The challenge for the formal learning system is to harness this insight to deliver more effective RPL processes, one where the perspectives of each side are acknowledged. It is the intention of this research study to develop a set of concrete recommendations to support the confidence and capability of the main actors within RPL. The research proposes that building capacity through collaborative networks will support assessors and RPL candidates and nurture capability for assessing non-formal and informal learning in practice.

1 Introduction

The focus of this study is to examine the values, beliefs and assumptions of the actors within Recognition of Prior Learning (RPL) in Ireland, to see if there are different perspectives, and to consider how to marry these perspectives through collaborative interventions to build trust and agency. RPL is a significant aspect of the current European policy climate on lifelong learning which has seen much restructuring and updating in recent years (Bologna, 1999, Bologna, 2001, Bologna, 2003, Bologna, 2007, Bologna, 2009, Council of the European Union, 2009, Crosier et al., 2012). In 2012, the Council of Europe recommended that all educational institutions have arrangements for RPL in place by 2018 (Council of the European Union, 2012). This has resulted in a revision of both national and local arrangements for RPL across the formal learning system (European Commission, 2010, Werquin, 2012). The push to increase participation in lifelong learning and to broaden access to a more diverse student population has resulted in RPL taking a central position within many reports and communiqués arising from the European Commission, UNESCO, OECD and CEDEFOP (Bologna, 1999, Bologna, 2001, Bologna, 2003, Bologna, 2007,

Bologna, 2009, Council of the European Union, 2009, Crosier et al., 2012, Werquin, 2010, UNESCO, 2012, CEDEFOP, 2015).

RPL allows for all forms of learning to be valued, regardless of how or where it was gained in the context of a destination award (Fiddler et al., 2006, Werquin, 2010, Duvekot et al., 2007). RPL allows for learning gained from activities in the workplace, professional training, or learning arising from voluntary activities contribute towards an award on the national framework of qualifications. In practice RPL is used to allow for non-standard entrants, for advanced entry or for credits for modules within programmes (NQAI, 2005, Higher Education and Training Awards Council, 2009).

2 Current situation

In spite of the drive in policy to provide for the recognition of all forms of learning within the formal learning system, in practice RPL is a complex brief to deliver:

1. RPL is used in different ways within institutions (Harris and Wihak, 2011);
2. Commonly, academic assessors experience doubt or hesitation about the assessment of non-formal and informal learning (Hewson, 2008);
3. The RPL candidate has to present their learning in a form appropriate for assessment; this task commonly presents as a complex brief in terms of writing, reflection and the selection of appropriate evidence (Conrad and Wardrop, 2010).

First, as stated above, RPL is used in various ways, for access or possibly for individual credits, institutions commonly make their own local arrangements for its provision, typically with a particular student cohort in mind (Harris, 2000). Such local arrangements are seen as necessary, and fit for purpose within that particular context. However, as a result, RPL provision takes on a diverse, almost patchwork effect and confidence around how RPL can be used, and capacity or confidence around its provision at a macro level is diluted amongst providers (Štastná, 2012, Hewson, 2008).

Secondly, the diverse and varied nature of the evidence and learning that is assessed through RPL raises doubts amongst academic assessors around its assessment. Questions arise around sufficiency and type of evidence, its currency and authenticity. Academic assessors commonly ask, how much evidence is enough and thus can harbour doubt as to what is sufficient (Hewson, 2008).

A third aspect related to the difficulties encountered within RPL stem from the context specific nature of non-formal and informal learning which the candidate for RPL will bring with him/her to the assessment process (Kawalilak and Wihak, 2013). It is a complex task to make such learning 'fit' within the formal learning system and this presents a major challenge for the candidate to present suitable forms of evidence for assessment (Conrad and Wardrop, 2010, Hamer, 2011, Hamer, 2012, Starr-Glass, 2012).

Therefore, cultural acceptance of RPL amongst relevant staff and candidates is a key factor cultivating conditions for pedagogic agency for the practice of RPL according to Cooper and Harris (2013). The literature suggests that once policy and procedures are in place, it is the acceptance and promotion of RPL by the academic staff which is key to its successful adoption within an institution (Leiste and Jensen, 2011, Kenniscentrum EVC, 2007).

The current research study is employing an innovative approach to supporting RPL provision by examining the values, beliefs and assumptions of both RPL assessors and candidates in order to explore in-depth the perspectives of both actors involved (Friesen, 2011). The overall aim of the

study is to develop insights and critical knowledge on RPL which will result in a more comprehensive understanding of the necessary conditions required to build trust and cultivate pedagogic agency for RPL in one Irish HE institution.

3 Methodology and Method

This research study is utilising a constructivist grounded theory approach to explore the values, beliefs and assumptions of academic assessors and candidates for RPL from their individual perspectives (Charmaz, 2006, Bryman, 2012). The research, which began in 2015, is being carried out within one Higher Education institution in Ireland.

3.1 Research design

Approaches and practices with RPL vary between institutions, and also within individual schools and faculty. With this in mind, designing an approach which was able to capture context rich data, particular to a given situation was necessary. The viewpoints of the academic assessor and past RPL candidates were of interest, and these each formed a data set. The individual perspectives of the participants; both held valuable insights, the nuances of which had to be understood on analysis. Therefore this qualitative study chose a semi-structured interview approach to provide some structure to support analysis while allowing the participant the freedom to freely express themselves in conversation (Bryman, 2012). Nvivo software was selected to support the initial analysis of this research still in progress. A full and rigorous analysis of the interview data is still to be undertaken. The interview questions were piloted with two academic assessors and two past RPL candidates, to afford the opportunity to amend the questions in light of the emerging responses (Bryman, 2012).

3.2 Sample selection of participants

Participants were selected on the basis of their having been involved previously in RPL. The sampling approach used random stratified sampling (Bryman, 2012) and involved interviewing 31 academic assessors in 2015, and 27 past candidates of RPL in 2016 using the same type of data collection questions (Figure 1). Participants were selected across a broad a range of disciplines; Business and Humanities, Science and Engineering, Art, and Maritime and programme levels, i.e. 6 to 9 on the Irish National Framework of Qualifications.

3.3 Procedure

For each interview, background information was initially provided before commencing with the questions. Ethical guidelines were strictly followed (Bryman, 2012). Face to face and telephone interviews were carried out by the researcher between March and June of 2015 and again in the spring of 2016.

1	What are your expectations of RPL?
2	In considering RPL what do you think are important values[1] to have?
3	Why do you say this?
4	What beliefs[2] do you hold about RPL?
5	Why are these beliefs important?
6	What do you assume[3] has to be in place before RPL can successfully be delivered?
7	Is there anything else that needs to be considered?

Note:

1. A value is that which is held as important and provides a framework as to how we live, think or act (Turner, 2004).
2. Beliefs are 'understandings, premises, or propositions about the world that are felt to be true' (Richardson, 1996, p. 196).
3. Assumptions are things we consider to be correct at a given time (Searby, 2009).

Figure 1: Interview questions

Each interview was recorded and fully transcribed and returned to each participant for respondent validation before initial data analysis commenced. The theoretical framework used to support analysis of the data is drawn from Ravitch and Riggan (2012), taking personal, topical and theoretical elements each informing the extraction of the data from the initial open coding step. The personal elements drawn from the researchers own perspectives within RPL provision, supporting the candidates primarily with case preparation. The topical aspects derive from the RPL literature on agency for RPL operation (Cooper and Harris, 2013) and the current policy push to have arrangements for RPL in place by 2018 (Council of the European Union, 2012, Department of Education & Skills, 2016, Quality and Qualifications Ireland, 2015). The theoretical element drew heavily on the work of English sociologist, Basil Bernstein whose concepts of the classification of knowledge and knowledge boundaries relate to the recognition of all forms of learning (Bernstein, 2000). In addition the concept of the totally pedagogised society relates to the operation of RPL, wherein populations are encouraged to repeatedly avail of upskilling and reskilling opportunities over a lifetime (Bernstein, 2000). Finally the field of recontextualisation is where knowledge is repositioned to facilitate entry to the formal learning system through assessment for RPL (Bernstein, 2000). Bernstein's theories are about language, its codes and rules, and how the structure of communication influences an individual's identity.

To date only an initial analysis has been carried out on the data. Transcripts were read and initial observations were recorded as memos. This was followed by initial open coding of the transcripts under each interview question. Similar codes were grouped together as themes. These are presented in Table 1 as the dominant codes.

4 Results

Presenting the findings to date, reveals an initial glimpse of the values, beliefs and assumptions of both sets of participants in RPL. Initial analysis suggests that the two main actors in RPL, the academic assessor and the RPL candidate take a different viewpoint, which is apparent in the values, beliefs and assumptions identified. Table 1 presents a summary of the findings arising from the analysis of the interview data to date and will be explained through the dominant codes that have emerged. At this point the sub-ordinate codes are still to be fully developed.

Academic Assessor of RPL material	Past RPL candidate
Values: Upholding of standards; non-judgemental	**Values:** Integrity**;** validity
Beliefs: Providing alternative pathways into education; the value of learning gained non-formally and informally	**Beliefs:** RPL is beneficial; RPL lightens the load
Assumptions: Procedures and protocol; advice	**Assumptions:** Systems and processes; validity of the case presented

Table 1: Summary findings showing the dominant codes of both academic assessor and past RPL candidate.

Visible within the data are tensions arising from the repositioning of every-day working knowledge into a form suitable for validation by the formal learning system (Starr-Glass, 2012). It is apparent that knowledge gained in the workplace and in context specific ways is inherently different to knowledge gained in formal settings (Kawalilak and Wihak, 2013).

The first set of results deal with the academic assessor and show that he/she looks first to maintain the academic standards. In addition they value holding a non-judgemental viewpoint in assessment. Initial analysis suggests that half of the initial coding arising from the interviews with the academic assessor are all themed under defending the standards of the institution. However, also visible within the data is that they are willing, where merited, to step towards supporting the candidate's RPL claim. This is revealed as the act of 'balancing,' and echoes similar findings arising from the research of Starr-Glass (2012). Balancing denotes that point when the academic assessor sees merit in the prior learning case, and they are willing to pause, hold judgement, and read on; to see how the candidate demonstrates in writing (normally) and through the evidence provided, what it is they know about a given set of learning outcomes/standards. In balancing, the assessor is reaching out from the traditional curriculum, delivery and assessment approaches towards the world of the workplace and community.

The second set of results also show that the RPL candidate possesses values, beliefs, and assumptions that strongly defend their reasons for applying. RPL is seen as a tool that can help ease the workload of the RPL candidate (Fiddler et al., 2006, NQAI, 2005). Another strong theme arising from the RPL candidate is the validity or the idea of integrity behind their application. The RPL candidate does not want to put themselves at any disadvantage by applying for RPL. The many comments highlighted the importance of honesty and integrity within the RPL application.

5 Conclusion and future work

The analysis of the data so far reveals two viewpoints operating within RPL. On the one hand the data arising from the interviews with the academic assessor suggests that they maintain the academic standards of programmes. All of the values, beliefs and assumptions are strongly aligned to protecting these academic standards. However, there is another aspect within the findings, that of the academic assessor reaching out towards or, trying to reach a balance between the formal learning system and that knowledge which is gained in non-formal and informal situations. This data provides useful insight, that where merited, an academic assessor will reach out to accomodate

the candidate and their prior learning claim (Kawalilak and Wihak, 2013, Starr-Glass, 2012). Academic assessor number 31 demonstrates this professional crux ;

> "I would see that I sit between both of those, supporter and gatekeeper."
>
> Academic assessor 31

The opinions of the RPL candidate tell an equally insightful story. The values, beliefs and assumptions of the past RPL candidate echo with the themes of integrity, and validity while acknowledging that RPL does lighten the workload within a programme. RPL is carefully chosen as an option when embarking on a programme, showing that the candidate is cautious not to put themselves at a disadvantage by opting for an RPL assessment:

> "You have the prior knowledge, and you really have to know what you are doing. If you are not up to speed with what you intend to get the RPL for you are setting yourself up for trouble, do you know what I mean?"
>
> RPL candidate 11

To summarise, the academic assessor will defend the standards strongly, while the RPL candidate strongly expresses their integrity and the validity of their case.

The challenge for the formal learning system is to deliver more effective RPL, one where the various perspectives are acknowledged. This research recommends actions which support building the confidence and capability of the main actors within RPL (Starr-Glass, 2012, Leiste and Jensen, 2011). Although this is research is still a work in progress, the findings so far suggest areas for consideration such as the provision of:

1. Readily accessible information on RPL for both candidates and academic assessors available at a local and national level
2. Opportunities to network when creating the RPL case for candidates, providing examples and scenarios to illustrate the outcomes of others
3. Practitioner networks for academic assessors where they can share expertise and raise questions with colleagues both locally and nationally
4. Training programmes for academic assessors to be made available both locally and nationally.

At the core of this is the concept of creating and maintaining collaborative networks which operate at both macro and local levels to support the actors within RPL (Goggin et al., 2015, Štastná, 2012). The values, beliefs and assumptions of the actors in RPL act to support their real-time evaluation of the pathways within RPL case preparation and assessment. The challenge will be to provide a range of fora to share practice and scenarios both at a local and national level in order to build trust and capability with RPL in all of its forms.

6 Acknowledgment

Thanks to both sets of participants for their support for my research and for their time and opinions which were given most generously.

7 References

Bernstein, B. 2000. *Pedagogy, sybolic control and identity; Theory research and critique*, Lanham, MD, Rowman & Littlefield.

Bologna 1999. *The Bologna Declaration of 19th June 1999. Joint Declaration of the European Ministers of Education*. European Union, Brussels, available at: www. bologna-berlin2003. de/pdf/bologna_declaration. pdf.

Bologna 2001. *Towards the European Higher Education Area: The Prague Communiqué*. Brussels: Council of the European Union

Bologna 2003. *Realising the European Higher Education Area: The Berlin Communiqué*. Brussels: Council of the European Union

Bologna 2007. *Towards the European Higher Education Area: Responding to Challenges in a Globalised World. The London Communiqué*. Brussels: Council of the European Union.

Bologna 2009. *The Bologna Process 2020: The European higher education area in the new decade. The Lewen and Louvain-la-Nevue Communiqué*. Brussels: Council of the European Union.

Bryman, A. 2012. *Social Research Methods*, Oxford, Oxford University Press.

CEDEFOP 2015. *European guidelines for the validation of non formal and informal learning*. Luxembourg: Publications Office, CEDEFOP no. 104.

Charmaz, K. 2006. *Constructing Grounded Theory: a Practical Guide through Qualitative Analysis*, London, Sage.

Conrad, D. & Wardrop, E. 2010. *Exploring the relationship of mentoring to learning in RPL practice*. Canadian Journal for the Study of Adult Education, 23, 1-22.

Cooper, L. & Harris, J. 2013. *Recognition of prior learning: exploring the 'knowledge question'*. International Journal of Lifelong Education, 32, 447-463.

Council Of The European Union 2009. *Notices from European Union Institutions and Bodies. Council conclusions of 12 May 2009 on a strategic framework for European cooperation in education and training ('ET2020')*. Brussels: Official Journal of the European Union.

Council Of The European Union 2012. *Council recommendation of 20 December 2012 on the validation of non-formal and informal learning. Official Journal of the European Union*. Brussels: Council of the European Union.

Crosier, D., Horvath, A., Kerpanova, V., Kockanova, D., Parveta, T., Dalferth, S. & Rauhvargers, A. 2012. *The European Higher Education Area in 2012: Bologna Process Implementation Report*. Brussels: European Commission, Education, Audiovisual and Culture Executive Agency.

Department of Education and Skills 2016. *Action Plan for Education 2016-2019*. Department of Education and Skills. Dublin.

Duvekot, R., Charraud, A.-M., Klarus, R., Schuur, K., Hoeij, K. V., Bórnávold, J., Konrad, J., Coughlan, D., Scanlon, G., Nilsen-Mohn, T., Halba, B., Gerster, A. C., Bonsema, P., Ferrari, L., Mancinelli, E., Paulusse, J., Walter, A., Cihaková, H., Stretti, M., Marinková, H., Alsma, E., Sanou, L., Krech, U. & Klenk, W. 2007. *Managing European diversity in lifelong learning; the many perspectives of the Valuation of Prior Learning in the European workplace*. Amsterdam: HAN University, Foundation EC-VPL & Hogeschool van Amsterdam.

European Comission 2010. *Recommendation on the promotion and validation of non-formal and informal learning*. Brussels: European Commission.

Fiddler, M., Marineau, C. & Whitaker, U. 2006. *Assessing Learning, Standards, Principles and Procedures*, Chicago: CAEL.

Friesen, N. 2011. Endword: *Reflections on research for an emergent field* In: J. Harris, M. Brier & C. Wihak (eds.) Researching the Recognition of Prior Learning; International Perspectives. Leicester: NIACE.

Goggin, D., Sheridan, I., O'Leary, P. & Cassidy, S. 2015. *A Current Overview of Recognition of Prior Learning in Irish Higher Education*. Dublin: National Forum for the Enhancement of Teaching & Learning in Higher Education.

Hamer, J. 2011. *Recognition of prior learning (RPL): can intersubjectivity and philospohy of recgnition support better equity outcomes?* Australian Journal of Adult Learning, 51, 90-109.

Hamer, J. 2012. *An Ontology of RPL: improving non-traditional learners' access to the recognition of prior learning through a philosophy of recognition*. Studies in Continuing Education, 34, 113-127.

Harris, J. 2000. *RPL: Power, pedagogy and possibility*, Pretoria, Human Sciences Research Council.

Harris, J. & Wihak, C. 2011. *Introduction and overview of chapters*. In: Harris, J., Brier, M. & Wihak, C. (eds.) Researching the Recognition of Prior Learning, International Perspectives. Leicester: NIACE.

Hewson, J. 2008. *RPL policy to practice: why the reticence of practitioners to engage?* In: Australian Vocational Education and Training Research Association (AVETRA) 11th Annual Conference, 3-4th April 2008 Adelaide.

Higher Education and Training Awards Council 2009. *Assessment and Standards; Implementing the National Framework of Qualifications and Applying the European Standards and Guidelines*. Dublin: Higher Education and Training Awards Council.

Kawalilak, C. & Wihak, C. 2013. *Adjusting the fulcrum: How prior learning is recognised and regarded in university adult education contexts*. College Quarterly, 16, 169-173.

Kennicentrum EVC 2007. *The covenant; a quality code for APL – identifying and accrediting a lifetime of learning*. Utrecht Kenniscentrum EVC.

Leiste, S. M. & Jensen, K. 2011. *Creating a positive prior learning assessment (PLA) experience: A step-by-step look at university PLA*. The International Review of Research in Open and Distance Learning, 12, 61-79.

NQAI 2005. *Principles and Operational Guidelines for the Recognition of Prior Learning in Further and Higher Education and Training*. Dublin: NQAI.

Quality and Qualifications Ireland 2015. *Policy Restatement: Policy and criteria for access, transfer and progression in relation to learners for providers of further and higher education and training*. Dublin: Quality and Qualifications Ireland.

Ravitch, S. & Riggan, M. 2012. Reason & Rigor, *How Conceptual Frameworks Guide Research*, Thousand Oaks, California, Sage.

Richardson, V. 1996. *The role of attitudes and beliefs in learning to teach*. In: Sikula, J. (ed.) Handbook of research on teacher education. New York: McMillan.

Starr-Glass, D. 2012. *Partial Alignment and Sustained Tension: Validity, Metaphor, and Prior Learning Assessment. PLA Inside Out: An International Journal on Theory*, Research and Practice in Prior Learning Assessment, 1.

Štasná, V. 2012. *EURASHE Seminar on Recognition of Prior Learning (RPL): Flexible Ties within Higher Education*. Prague: EURASHE.

Turner, M. 2004. *Values & beliefs in mentoring*. Coach the Coach [Online]. Available: http://www.mentoringforchange.co.uk/pdf/CtC%20-%20Values.pdf (accessed on 29th October 2016).

UNESCO 2012. *UNESCO Guidelines for the Recognition, Validation and Accreditation of Non-formal and Informal Learning*. Hamburg: UNESCO Institute for Lifelong Learning.

Werquin, P. 2010. *Recognition of Non-Formal and Informal Learning; Outcomes, Policies and Practices*. Paris: OECD.

Werquin, P. 2012. *The missing link to connect education and employment: recognition of non-formal and informal learning outcomes*. Journal of Education and Work, 25, 259-278.

Call for Papers
for next Collaborative European Research Conference (CERC 2017)

Celebrating Excellence in Research

The multidisciplinary Collaborative European Research Conference (CERC) is an annual event that takes place since 2011 when it was initiated by University partners across Europe. It bring together researchers from a wide range of disciplines in order to foster knowledge transfer, inter-disciplinary exchange and collaboration.

We invite researchers from all third level academic institutions, whose research topics are related to applied computer science, to submit their papers and to attend the conference. We especially welcome postgraduate researchers who wish to present new research topics.

In previous years contributions came from biology, computer science, bioinformatics, business information systems, marketing, it-security, civil engineering, education, psychology, multimedia, art, etc. Contributing students came from Ireland, Germany, England, Wales, Northern Ireland, Norway, France, USA, etc.

CERC always had two main aims :

- to provide showcases for early stage researchers
- to celebrate excellence in research

CERC will provide an opportunity for students to engage with researchers as well as with practitioners by creating a forum for the collaborative discussion of ideas.

Date and Venue :
CERC 2017 will be hosted by the University of Applied Sciences Karlsruhe (HSKA) Germany and the Steinbeis Stiftung Karlsruhe, September 22nd and 23rd 2017.

Submission Details :

Author Registration :	150 €
Guest Registration :	150 €

The registration fee includes a conference programme including refreshments, lunch and proceedings as well as a social event. Accommodation must be pre-booked separately. Recommended accommodations and negotiated conference rates will be made available on the conference webpage.

The paper submission will consist in a two stage process. In the first stage short papers with a max. of 3 pages will be submitted. These short papers will be peer reviewed. During the conference selected contributions will be invited to submit a full paper (10 pages). Both types of accepted contributions will be published in digital proceedings with an ISSN number. A prize for the best paper will be awarded during the closing ceremony on the evening of the 22nd of September 2017.

The link to the paper submission site as well as registration will be made available on the webpage : www.cerc-conference.eu .

Schedule

A preliminary programme and social event will be available at the conference website. Please note that this is subject to change.

Important dates :

Short-paper submissions due	31st of March
Author notification	30th of June
Submission of camera ready papers	30th of July
Registration until	31st of August
Conference	22nd, 23rd of September

Conference Awards

Best Paper / Presentation Award

presented to :

Lawlor, Walsh

Metamorphosis: Changing the shape of genomic data using Kafka

Call for Papers for CERC 2017

see page 238

or CERC Website : www.cerc-conference.eu

ISSN : 2220 - 4164