

# **Proceedings of the International Conference on Analytics Driven Solutions**

**IBM Centre for Business Analytics and  
Performance**

**at the  
Telfer School of Management  
University of Ottawa  
Canada**

**29-30 September 2014**



**Edited by**

**Dr Eduardo Rodriguez and Dr Greg Richards**

**acpi**

A conference managed by ACPI, UK



**The Proceedings of  
The  
International Conference on  
Analytics Driven Solutions  
ICAS 2014**

**IBM Centre for Business  
Analytics and Performance  
Located at the Telfer School of Management  
University of Ottawa  
Canada  
29-30 September 2014**

Edited by  
Dr Eduardo Rodriguez and Dr Greg Richards  
Telfer School of Management  
University of Ottawa and IQAnalytics Inc.  
Canada

Copyright The Authors, 2014. All Rights Reserved.

No reproduction, copy or transmission may be made without written permission from the individual authors.

Papers have been double-blind peer reviewed before final submission to the conference. Initially, paper abstracts were read and selected by the conference panel for submission as possible papers for the conference.

Many thanks to the reviewers who helped ensure the quality of the full papers.

These Conference Proceedings have been submitted to Thomson ISI for indexing. Please note that the process of indexing can take up to a year to complete.

Further copies of this book and previous year's proceedings can be purchased from <http://academic-bookshop.com>

E-Book ISBN: 978-1-910309-50-6

Book version ISBN: 978-1-910309-48-3

CD Version ISBN: 978-1-910309-49-0

Published by Academic Conferences and Publishing International Limited

Reading

UK

44-118-972-4148

[www.academic-publishing.org](http://www.academic-publishing.org)

## Contents

Paper Title	Author(s)	Page No.
<b>Preface</b>		iii
<b>Committee</b>		iv
<b>Biographies</b>		v
<b>Research papers</b>		
Context Aware Air Traffic Management System	Nihad Al-Juboori and Saif Al-Sultan	1
Post Implementation Of ERP System	Imran Batada and Asmita Rahman	12
An Analysis of Canadian Armed Forces Historical Release Rates by Location	Din Begovic	20
Importance, Use and Organizational Assignment of Web Analytics for Social Economy Organizations in Austria – An initial empirical study	Claudia Brauer, Stefan Bauer and Julia Hauptmann	30
Foundations of Big Data and Analytics in Higher Education	Ben Daniel and Russell Butson	39
Data, Information and Intelligence	Scott Erickson and Helen Rothberg	48
Building an Undergraduate Program in Business Analytics	Scott Erickson, William Tastle and Danielle Puleo	55
Using Entropy for Healthcare Analytics and Risk Management in Influenza Vaccination Programs	Eugene Levner, Sharon Hovav, Hanan Tell, Alexander Ptuskin, Avi Herbon	63
Google Analytics as a Prosumption Tool for Web Analytics	Joanna Palonka, Marcin Lora, Tomasz Oziębło	73
Analytics in Context: Modelling in a Regulatory Environment	Albert Simard	82
Online Community Projects in Lithuania: Cyber Security Perspective	Aelita Skaržauskienė, Darius Štītis, Monika Mačiulienė, Gintarė Paražinskaitė and Agnė Tvaronavičienė	93
Pocket Data Analytics: An Introduction to Practical Applications of Course Analytics in Online Learning	John Vivolo	102
<b>PHD Research Paper</b>		111
Exploring Analytics in Health Information Delivery to Acute Health Care in Australia	Chandana Unnithan, Katerina Andronis	113
Information Visualisation and Knowledge Re-construction of RFID Technology Translation in Australian hospitals	Chandana Unnithan, Arthur Tatnall , Stephen Burgess	122
<b>Masters Research Paper</b>		133
Finding Patterns in Medical Ward Data Using Rough Sets	Anwar Alenezi and Julia Johnson	135
<b>Non Academic Paper</b>		145
Statistical Validation of a new Python-Based Military Workforce Simulation Model	Stephen Okazawa, Patricia Moorhead, Abe Jesion and Stan Isbrandt	147
Development of a Naval Reserve Training and Career Progression Analysis Application	Michelle Straver	156
Detailed Maintenance Planning for Military Systems with Random Lead Times and Cannibalization	Rui Zhang and Ahmed Ghanmi	166

<b>Paper Title</b>	<b>Author(s)</b>	<b>Page No.</b>
<b>Work In Progress paper</b>		175
Knowledge-Based Cause-Effect Analysis Enriched by Generating Multi-Layered DSS Models	Axel Benjamins	177
A Simulation Study of Operational Availability of Canadian Armed Forces Equipment under Different Maintenance Regimes	Abdeslem Boukhtouta and Ahmed Ghanmi	183

## Preface

## Conference Committee

### Conference Executive

**Dr Eduardo Rodriguez**, IQAnalytics Inc., Canada

**Dr Greg Richards**, Telfer School of Management, University of Ottawa, Ottawa, Canada

### Mini track chairs

**Dr Dan Shoemaker**, Centre for Cyber Security and Intelligence Studies, University of Detroit Mercy, USA

### Committee Members

The conference programme committee consists of key people in the analytics domain from a variety of different communities; the list includes leading academics, researchers, and practitioners from around the world. The following people have confirmed their participation:

Pavel Albores (Aston Business School, UK); Prof Abdelnaser Ali (Universiti Sains Malaysia, Malaysia); Jom Aw (Ottawa University, Canada); Alejandro Balbas (University Carlos III, Spain); Beatriz Balbas (Universidad Castilla La Mancha, Spain); Irma Becerra (Florida International University, USA); Dr Samarjeet Borah (Sikkim Manipal University, India); Fernando Camacho (Waterloo University, Canada); Sven Carlsson (Lund University School of Economics and Management, Sweden); Dr Juan-Gabriel Cegarra-Navarro (Universidad Politécnica de Cartagena, Spain); Dr Sunil Choenni (Ministry of Security & Justice, The Netherlands); Dr Tom Cockburn (Center for Dynamic Leadership Models in Global Business, New Zealand); Jonathan Crook (University of Edinburgh, UK); Kimis Dalkir (McGill University, Montreal, Canada); Babak Daneshvar Rouyendegh (Atilim University, Turkey); Dr Strasunskas Darijus (Applied Science, Norway); Dr Sofia Daskou (Hellenic American University, Greece); Dr Yan Qing Duan (Luton Business School, University of Luton, UK); Prof John Edwards (Aston Business School, UK); Prof Ali Emrouznejad (Aston Business School, UK); Dr Barbara Endicott-Popovsky (University of Washington Seattle, USA); Dr Scott Erickson (Ithaca College, New York, USA); Dr Turan Erman Erkan (Atilim University, Turkey); Prof Alptekin Erkollar (Halic University, Turkey); Jorge Ferreira (Nova University of Lisbon, Portugal); Elisabeth Frisk (Chalmers University of Technology, Sweden); Alexeis Garcia-Perez (Coventry University, UK); Dr Roya Gholami (Aston Business School, UK); Asst Prof Dr Poyraz Gurson (Atilim University, Ankara, Turkey); Joey Jansen van Vuuren (CSIR, South Africa); Dr Paul Jones (Plymouth University, UK); Selvi Kannan (Victoria University, Melbourne, Australia); Prof Jesuk Ko (Gwangju University, Korea); R. Prasad Kodaypak (ATT, USA); Prasad Kodypak (Researcher ATT, USA); Dr Kostas Kolomvatsos (Dept of Informatics and Telecommunications, University of Athens, Greece); Dr Jussi Koskinen (University of Jyväskylä, Finland); Dr Iwona Lapunka (Opole University of Technology, Poland); Alvaro Lecompte (Inter American University of Puerto Rico, Puerto Rico); Allena Leonard (Cybernetic Society, Canada); Jun Li (Communications Research Centre, Canada); Jay Liebowitz (UMUC, USA); Dr Paolo Magrassi (Alphatune, Italy); Mario Ernesto Martinez (Universidad La Sabana, Colombia); Rubi Mejia (EAFIT University, Colombia); Vilma Mesa (University Michigan, USA); Rene Meziat (Universidad del Rosario, Colombia); Delfin Montuno (IEEE and Carleton University, Canada); Manuel Morales (University Montreal, Canada); Mark Morreale (SAS, Canada); Dr Balsam Mustafa (University Malaysia Pahang, Malaysia); Prof Collins C. Ngwakwe (University of Limpopo, South Africa); Jian-Yun Nie (University Montreal, Canada); Dr Birgit Oberer (Kadir Has University, Turkey); Dr Brian O'Flaherty (University College Cork, Ireland); Asst Prof Sharada Patagupi (IETE, Bangalore, India); Dr Iwona Pisz (Opole University of Technology, Poland); Dr/Prof Zhan Qinglong (Tianjin University of Technology and Education, China); Prof Thurasamy Ramayah (Universiti Sains Malaysia, Malaysia); Dr German Ramirez (Laureate International Universities, USA); Dr Gabriel Ramirez (Universidad del Rosario, Colombia); Dr Jose Manuel Restrepo (CESA School of Business, Columbia); Greg Richards (University of Ottawa, Canada); Martyn Roberts (University of Plymouth, UK); Prof Julie Ryan (George Washington University, USA); Dr. Jatinderkumar R. Saini (Narmada College of Computer Application, Zadeshwar, Bharuch, Gujarat, India); Prof Abdel-Badeeh Salem (Ain Shams University, Egypt); Dr Elena Serova (St. Petersburg State University of Economics, Russian Federation); Yilun Shang (Singapore University of Technology and Design, Singapore); Dr Dan Shoemaker (University of Detroit Mercy, Detroit, USA); Dr. Albert Simard (Defence R&D Canada, Canada); Peter Smith (CDLMGB Faculty, Canada); Prof. dr. Marta-Christina Suci (Academy of Economic Studies Bucharest, Romania); Prof Reima Suomi (University of Turku, Finland); Pedro Szekely (University of South California, USA); Prof João Tavares (Faculdade de Engenharia da Universidade do Porto, Portugal); Dr Eduardo Tomé (Universidade Lusitana, Portugal, Portugal); Dr Mueen Uddin (Asia Pacific University KL Malaysia, Malaysia); Chandana Unnithan, Deakin University, Melbourne, Australia; Carlos Videla (Mount Royal University, Canada); Professor Jose Maria Viedma (Polytechnic University of Catalonia, Spain); Russell Walker (Northwestern University, USA); Prof Stanislaw Wrycza (University of Gdansk, Department of Business Informatics, Poland)



## Biographies

### Conference Chair



**Dr Eduardo Rodriguez** teaches in the Telfer School of Management at Ottawa University. He is also a Senior Associate of the Center for Dynamic Leadership Models in Global Business at The Leadership Alliance Inc. and Principal of IQAnalytics Inc., a research centre and consulting firm. He has extensive experience in knowledge and risk management, and in analytics in the healthcare, insurance and banking industries. He has been a knowledge management advisor to EDC( Export Development Canada) in Ottawa, Regional Director of PRMIA (Professional Risk Managers International Association) in Ottawa, Director of Strategic Intelligence UNAD Colombia, Professor at the Andes University and CESA, Quantitative Analyst for EDC, and Senior Executive in banks and insurance companies in Colombia. Eduardo is a Mathematician and holds an MBA and MSc. in Mathematics. He is certified by the Advanced Management Program at McGill University and has a PhD from Aston Business School in the UK. His main research interest is in the field of Knowledge Management applied to Enterprise Risk Management. He is author of four books in analytics, a reviewer of four academic journals in the analytics field, and has publications in peer-reviewed journals and conferences. He is an international lecturer in the field of analytics. Currently, he is leading the permanent Think-Tank in Analytics in Ottawa with world-wide scope.

### Programme Chair

**Dr Greg Richards** conducts research into performance management practices and outcomes in public and private sector organizations. He coordinates, in collaboration with Professor Swee Goh, the Interis Public Sector Performance Management research cluster. The focus of this cluster is on the approaches used to stimulate high performance in public sector organizations. In addition, he currently oversees the development of research and curriculum building initiatives for the newly created IBM Centre for Performance Management. The focus of the Centre is on the application of Business Analytics methods and technologies in enabling organizational high performance. Professor Richards has taught Organizational Behaviour, Human Resource Management, Organization Design and Change Management. He currently teaches courses on Corporate Performance Management and Business Analytics in the MBA and EMBA programs. In addition, he teaches the Management Consulting course and co-coordinates the MBA Consulting project. Prior to his academic appointment, Greg worked at Transport Canada and at Consulting and Audit Canada as well as with Cognos Incorporated. He also has over 20 years of consulting experience and is a Certified Management Consultant.



### Mini Track Chair



**Dr Dan Shoemaker** is the author of four text books in the field of cybersecurity. He is also the Co-Chair of Workforce Training and Education (WET) for the Department of Homeland Security's Software and Supply Chain Assurance initiative. He is also a consultant for the US Department of Defence (DoD) Trusted System Initiative (TSI) specializing in international engagement. Dan writes and speaks frequently on cyber security topics and he is a full professor at the University of Detroit Mercy for his day-job

## **Biographies of Presenting Authors**

**Anwar Fahat Alenezi** is studying for a Master's degree in Computational Sciences at Laurentian University in Sudbury, Ontario, Canada. He has a Bachelor of Education degree from the Faculty of Teachers in Computer Sciences at Northern Border University, Arar, Saudi Arabia.

**Dr. Nihad Al-Juboori** received his PhD degree in Computer Engineering from Al-Nahrain University, Iraq. He is currently the Dean of Students Affairs at the Applied Science University, Kingdom of Bahrain. His research interests include Genetic Algorithms, Image Processing and intelligent transportation systems.

**Din Begovic** has been working as a defence scientist at the Department of National Defence since 2010. His primary duties include Canadian Armed Forces workforce modeling and analysis. In addition to working for the federal government, Din is at the thesis defence stage of his PhD in Operations Management at Carleton University.

**Axel Benjamins** holds a Master of Science in Information Systems (2009). He has been a research assistant (research and teaching) in the department of Management Support Systems at the University of Osnabrueck since 2012. His research topics are knowledge-based systems and decision support.

**Abdeslem Boukhtouta** is a defence scientist with the Defence Research and Development Canada, Centre of Operational Research & Analysis. He earned a Ph.D. degree from the Decision System and Operations Department, Laval University, and an MS degree in Civil Engineering from Sherbrooke University, both from Quebec, Canada. His research interests include operations research and stochastic dynamic programming related fields.

**Prof. Dr. Claudia Brauer** is Professor for Business & Management at the MCI Innsbruck. For more than 12 years Mrs. Brauer has been researching in the fields of Online Marketing, Social Media Marketing and Mobile Marketing and the performance measurement of Internet-based business models (Web Analytics, Social Media Analytics, Mobile Analytics).

**Dr. Ben Daniel** is a Senior Lecturer at the University of Otago, New Zealand. He lectures on Educational Technologies, Research Methodologies and Statistics to academics and postgraduate students. He currently heads the Educational Technology group at the Higher Education Development Centre, and leads the research programme on Big Data in higher education.

**Dr G. Scott Erickson** is Professor of Marketing in the School of Business at Ithaca College, Ithaca, NY where he has also served as Department Chair and Interim Associate Dean. He holds a PhD from Lehigh University, Masters degrees from Thunderbird and SMU, and a BA from Haverford College.

**Sonia Latchman** leads the Workforce Modelling and Analysis (WFMA) team within Director General Military Personnel Research and Analysis (DGMPPRA). Research interests include manpower and personnel modelling, attrition behaviour and forecasting, and strategic HR analysis. She holds a BSc in Mechanical Engineering from the University of Manitoba.

**Eugene Levner** holds a PhD degree in Computer and Systems Sciences from the USSR Academy of Sciences, Moscow (1972). He is Full Professor at Ashkelon Academic College and Holon Institute of Technology (Israel). His research interests are *Design of computer algorithms, Optimization, and OR in Healthcare*. He is author of 130 papers and 4 books.

**Stephen Okazawa** is a defence scientist in Defence Research and Development Canada. His research focusses on the development of discrete event simulation methods to inform military decision making, particularly in the area of military workforce analysis and forecasting. He holds a Master's degree in Electro-Mechanical Engineering from the University of British Columbia, Canada.

**Dr Joanna Palonka** is Assistant Professor in the Faculty of Informatics and Communication at the University of Economics in Katowice. She lectures and does research on Information Systems Design and Internet Communication Management. She takes part in the National Research Centre Project: Information System Prosumption for Support of Business Organization Management.

**Dr. Albert J. Simard** led the development of two award-winning national forest-fire information systems and strategic plans for two global disaster management information networks for the Canadian Government. He has led projects and developed strategies for a number of federal departments, including an NRCan access to knowledge policy, a DRDC knowledge services framework, and a CFIA modeling framework.

**Dr Aelita Skaržauskienė** is Director of the Business and Media School at Mykolas Romeris University, Vilnius, Lithuania, together with Middlesex University, UK. In her work she applies knowledge of management and modern leadership-correlated

disciplines such as Business dynamics, Systems thinking, Chaos and Complexity theories. Previously, she was the coach in the Self-Managing team's project in the European Parliament together with the DEMOS Group Belgium.

**Michelle Straver** obtained her BSc at Queen's University in 2000, and her MASc at the University of British Columbia in 2002, both in the field of Mechanical Engineering. She has been with Defence Research and Development Canada since 2002. Her recent research interests have included attrition analysis and forecasting, training systems, and occupation structure sustainability.

**Chandana Unnithan** is an academic at Deakin University, School of Information and Business Analytics, Australia. Her career spans a decade in higher education and another decade in the ICT sector with multinationals such as IBM. She is a published author in information systems, electronic governance, health informatics, and social media. Her current focus is in visual analytics.

**John Vivolo** is the Director of Online and Virtual Learning for New York University. He has worked in online learning for over ten years partnering with faculty to devise a set of pedagogical and technical best practices. John has a Master's in English from the City University of New York and is Sloan C certified.

**Dr Rui Zhang** obtained his PhD in Operations Management from the DeGroote School of Business, McMaster University Canada. Currently, he is a NSERC-supported VF working at DRDC-CORA. His research is mainly about detailed operational planning with either deterministic or stochastic situations. His papers have been published in journals such as Operations Research and Journal of Scheduling.



# Context Aware Air Traffic Management System

Nihad Al-Juboori and Saif Al-Sultan

Applied Science University, Manama, Bahrain

[Dr.nihad@asu.edu.bh](mailto:Dr.nihad@asu.edu.bh)

[s.j.alsultan@ieee.org](mailto:s.j.alsultan@ieee.org)

**Abstract:** The main goal of modern Air Traffic management systems is safety of the flights, efficient and proficient operation under the controller's responsibility. However, these systems have to satisfy user preferences and convenience to the maximum extent as possible. Providing safety requires a reliable detection and an effective resolution of impending intersections of airplane trajectories, which is referred to as conflict resolution. In this paper, a context-aware architecture for solving the problem of conflict resolution has been proposed. The architecture is divided into three stages which are sensing, reasoning and acting stage. In the reasoning stage, an algorithm for generating resolved paths has been introduced. Genetic algorithms have been utilized to generate safe paths that optimize fuel consumption and time delay. Two ways are suggested to solve this problem by either 2D maneuvering depending on the heading angle as a control parameter or 3D maneuvering depending on both heading angle and flight path angle. Simulation results showed the validity and the effectiveness of the proposed algorithm in terms of satisfying user preferences by providing a smooth path. In addition, the algorithm was able to propose resolved paths that maintain g-load, have minimal fuel consumption and have less time delay.

**Keywords:** Air Traffic Management Systems, Context Aware Systems, Genetic Algorithms, Conflict Resolution, Fuel Consumption

---

## 1. Introduction

The main goal of modern Air Traffic Management systems (ATM) is safety of the flights, efficient and proficient operation under the controller's responsibility. However, these systems have to satisfy user preferences and convenience to the maximum extent as possible. Providing safety requires a reliable detection and an effective resolution of impending intersections of airplane trajectories, which is referred to as conflict resolution. The way to avoid airplanes collisions is by detecting whether airplanes will be colliding in advance. Hence, conflict detection could be defined as the process of prediction future airplanes trajectories depending on their positions, height, speed, and their heading angles (Vela et al., 2009), in order to check if airplanes loss their safety separation distances which is determined as 5 Nautical Miles laterally and 1000 feet vertically (Kuchar and Yang, 2000). Detecting the conflict is not the only target of an ATM; a resolution schema has to be provided in order to avoid the conflict.

Context-aware systems are those systems that are able to adapt their behavior according to the current environment; they have the ability to sense, reason and react to the current environmental context. Any context-aware system has to have three sub-systems which are sensing, reasoning and acting subsystem (Baldauf et al., 2007). Utilizing the concept context-awareness in air traffic management systems will lead to enhance their capabilities in terms of detecting and resolving the conflict without any interaction from human. Several conflict resolution algorithms have been proposed in the literature to solve the problem of conflict. Conflict resolution algorithms can be defined as the process of determining the action that should be carried out by airplanes in order to prevent the collision (Emami and Derakhshan, 2012). These algorithms differs from each other by many ways, such as the prediction of the future resolved trajectories, limits and constraints according to user preferences that affect the resolved path or trajectory, the amount of communication needed between air planes that are taking part in the conflict duration.

There are six phases that the airplane should undergo during the process of flying which are: flight plan, taxiing, take off, cruising, decent and landing (Khedkar and Kumar, 2011). In this paper, a context-aware air traffic management system architecture is introduced to solve the problem of airplanes conflict during cruising (en-route) phase. The architecture is based on the five-layered conceptual model (Dey et al., 2001), and is divided into three stages which are sensing, reasoning and application stage. In the sensing stage, the system collects information about the conflict. In the reasoning stage, a novel Genetic algorithm has been proposed to generate the resolved path in the case of conflict performing both 2D and 3D maneuvering. The application stage is responsible for presenting the resolved path to the pilot or for utilizing the autopilot to move the airplane according to the generated path. The proposed algorithm is able to generate paths that satisfy several criteria, such as optimizing the fuel consumption, minimizing the time delay and maintaining the g-load.

The rest of this paper is organized as follows. Section 2, discusses the work that have been carried out in this area. An overview of the conflict resolution process is presented in section 3. Section 4 introduces the proposed air traffic management context-aware architecture. In section 5, the proposed Genetic conflict resolution algorithm is illustrated. System evaluation and results is shown in section 6, and the conclusion is given in section 7.

## **2. Related Work**

Several researchers have tried to solve the problem of airplanes conflict by proposing different conflict resolution algorithms. A summary for the main work that has been done in this field is given below.

In (Vela et al., 2012), a method for reducing the conflict task loads (reducing the number of commands needed to resolve the conflict) was proposed. The algorithm proposes changing the heading angle and the airplanes speed for solving the conflict. It, meanwhile, minimizes the fuel consumption by two cost functions which are fuel consumption due to speed change and fuel consumption due to additional distance coming from heading angle changes. The authors also propose a function to minimize the number of maneuver required by each air plane as a penalty function for the cost. However, the algorithm solves the conflict by selecting a 2D resolution and neglecting the flight path angle which will lead to consumption in fuel in case where the airplane tends to change freely in 3D.

A Genetic algorithm has been proposed by (Zhang et al., 2014) to solve the problem of conflict resolution. The authors suggest changing only the airplane's heading angle which means solving the conflict in 2D maneuvering. In order to reduce the time delay, the algorithm tries to change the perpendicular component of the speed with the heading angle. The functionality of the proposed chromosome is to generate a new value for the heading angle which varies from -35 to 35 degree. The optimization of the fuel is suggested to be proportional to the shortest path neglecting the aerodynamic equations that directly affect the fuel consumption. The fitness function in the proposed algorithm depends only on the distance travelled without considering other factors. However, in our proposed algorithm, we have taken into account several factors regarding the optimization, such as fuel consumption, g-load, etc. In addition, our proposed chromosome provides the rate of change in heading and flight path angles, thus supporting both 2D and 3D.

In (Peyronne et al., 2010), the authors depends on the B-splines for providing smooth air planes trajectories. The conflict was defined by dividing the air space into small squares equals the half distance of the safety operations. They try to optimize the resolved path by taking into consideration only the distance without considering the time delay and the fuel consumption. Yet, another algorithm for conflict resolution has been introduced (Gao et al., 2012). They argued that there are many suggested resolved path for each air plane. Hence, each airplane will find its own path independently from each other. Then the conflicted airplanes will cooperate for selecting a new path for all of them depending on their suggested paths. This cooperation will be done depending on a penalty function called system efficiency, which normalizes the summation of each airplanes normalized paths. However, their suggested conflict resolution mechanism will be in 2D maneuvering only without discussing the factors that might affect the fuel consumptions.

The authors in (Malaek et al., 2011), have proposed a Genetic algorithm for conflict resolution. It was argued that the solution of airplanes' conflict can be solved by free flight to overcome the ATM limitations that appear in the incompetent use of air space and the centralized ATM management. The proposed algorithm suggests to divide the airspace into small squares referred to as web, the size of each part of the web depends on either trial and error or on user experience. The cost function that identifies the suitable path should depend on how the airplanes are involved in the conflict, thrust required for such a maneuver, time delay in the new path and the amount of fuel consumption. The proposed chromosome depends on the amount of deviation from the original path, the time required for decision and the load factor that each airplane should maintain through its maneuvers. However, the algorithm works for 2D maneuvers only.

Another Genetic algorithm has been introduced by (Karr et al., 2009). The maneuver was defined as a template or a pattern contains different ways for path modification. The resolved path will be generated by Genetic algorithm depending on many constraints such as the new pattern route that will solve the conflict, the time and the speed. The constraints can be extended to include any relevant criteria if the algorithm is going to be applied in real life. The resolved path will optimize the minimum path from the original one, fuel

consumption, lateral offset distance and the distance travelled in the offset path. However, the proposed algorithm takes into account the 2D maneuvers only. In (Chaloulos et al., 2010), the aerodynamic model for the airplane has been described. Meanwhile, the load factor (g-load) has not been considered as one of the resolved path constraints. However, our proposed work takes into account the load factor during generating the resolved path. In (Cobano et al., 2012), a conflict detection and resolution algorithms was presented. The algorithm suggests dividing the air space into zones and solving the conflict by changing the speed of the airplanes without changing their direction. Such a scheme is acceptable if the optimizations on the time delay and fuel consumption are not necessary, which are taken into account in our proposed algorithm.

In this paper, a context-aware air traffic management system that is able to solve the problem of airplanes conflict during en-route phase is proposed. An air traffic management system architecture based on the concept of context awareness has been introduced. The architecture is divided into three phases which are sensing, reasoning and acting subsystems. In the reasoning phase, a novel Genetic algorithm for conflict resolution in terms of 2D and 3D maneuvering is presented. To solve the problem of conflict, the algorithm decodes two control parameters which are the flight path angle and the heading angle. In the 2D resolution, the system will provide a path for the conflicted airplanes using only the cross range (lateral) and downrange (longitudinal) directions. Hence, the control parameter will be the flight path angle only because of the assumption that there is no need to change its height (as is obviously known the heading angle is responsible for the change of airplane height). While, in the 3D resolution, the algorithm will provide a path for the conflicted airplanes using all possible directions, which are lateral, longitudinal and vertical. Thus, both of the angles mentioned above have been utilized. During generating the resolved path, different parameters have been taken into account, such as fuel consumption, g-load, thrust validity, time delay and maintaining the required separation between airplanes.

### 3. Overview of Conflict Resolution

#### 3.1 Point Mass Airplane Model

The point mass aircraft model captures the dynamical effects encountered in civil aviation aircraft. The point mass model assumes that the aircraft thrust is directed along the velocity vector, the aircraft always performs coordinated maneuvers and a flat, non-rotating earth (Menon et al., 1999),(Blakelock, 1991). The following Equations describe aircraft flight::

$$\dot{V}_i = \frac{(T_i - D_i)}{m_i} - g \sin \gamma_i \quad \dots (1)$$

$$\dot{\gamma} = \frac{g}{V_i} \left( \frac{L_i \cos \phi_i}{g m_i} - \cos \gamma \right) \quad \dots (2)$$

$$\dot{\chi} = \frac{L_i \sin \phi_i}{m_i V_i \cos \gamma} \quad \dots (3)$$

$$\dot{m} = -Q_i \quad \dots (4)$$

$$\dot{x}_i = V_i \cos \gamma_i \cos \chi_i \quad \dots (5)$$

$$\dot{y}_i = V_i \cos \gamma_i \sin \chi_i \quad \dots (6)$$

$$\dot{h}_i = V_i \sin \gamma_i \quad \dots (7)$$

with  $i = 1, 2, 3, \dots, n$  being the aircraft under consideration. In Equations(1) - (7),  $V$  is the ground speed, which is assumed to be equal to air speed in this paper.  $T$  is the air craft engine thrust,  $D$  is the drag,  $m$  is the aircraft mass,  $g$  is the acceleration due to the gravity,  $\gamma$  is the flight path angle,  $L$  is the vehicle lift,  $\Phi$  is the bank angle,  $\chi$  is the aircraft heading angle,  $x$  is the down range,  $y$  is the cross range,  $h$  is the altitude and  $Q$  is the feul flow rate.

### 3.2 Path Formulation

As discussed earlier, our proposed algorithm works in the en-route phase. The en-route phase is usually described by a set of way points that will direct the pilot through the period of aviation. The system will divide the controlled path into segments referred to as Delta, each of which has the same time interval. Hence, the control process will focus on each Delta to insure that no conflict occurs. The system will recalculate all airplanes trajectories at the beginning of each Delta. There calculation will be carried out not for the upcoming Delta, but for the one next to it. The time duration for each Delta is assumed to be 300 seconds.

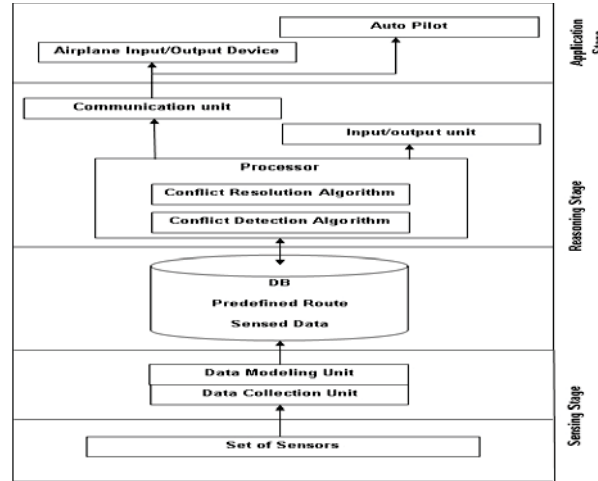
### 3.3 Conflict Detection and Resolution Algorithms

The conflict detection function has the responsibility to examine each Delta for each airplane under control, to insure that no conflict occurs during airplane's flight. The input for this function is the full trajectory for each airplane even though it has been modified by the system. However, the system has the ability to resolve the conflict even if it occurs by its previous resolution. At this stage, the output of the conflict detection algorithm which is airplanes Delta paths that are in conflict will be used as an input for the conflict resolution algorithm. Two parameters are responsible for the control of the airplanes during their flight, which are the flight path angle and the heading angle. The resolution algorithm will solve the conflict between air planes by performing suitable maneuvers that optimize the time and the fuel consumption.

## 4. Context-Aware-Based Air Traffic Management System Architecture

As mentioned earlier, there are three main sub-systems in any context-aware system, which are sensing, reasoning and action subsystems. Our proposed architecture is designed for the air traffic management system's ground station, and is divided into three main stages, sensing, reasoning and application stage. Presenting the generated path to the pilot or moving the vehicle according the generated route utilizing the auto pilot is the responsibility of the application stage depending on the output of the reasoning stage which in turn relies on the sensing stage. Thus, the process of detecting and resolving the conflict is a self-organizing process.

As shown in Figure 1, the architecture is divided into three stages as follows:



**Figure 1:** Context Aware Air Traffic Management System Architecture

- Sensing Stage: This stage is responsible for collecting the required data for detecting and resolving the conflict, it consists of two layers as follows:
  - Sensing layer: This layer comprises a set of sensors for collecting the data. Sensors might be physical (i.e. speed sensor), logical (i.e. ground station) or virtual (i.e.combined), (Indulska and Sutton, 2003).
  - Knowledge Representation layer: This layer consists of data collection and data modeling units. The former, is responsible for managing the sensors and separate their low-level details from the application. While, the latter, is responsible for transferring and storing the raw data gathered by sensors into a machine executable form using one of the available modeling methods. (modeling the data is out of the scope of this paper).



- Reasoning stage: This stage is responsible for detecting and resolving the conflict, it comprises the following components:
  - Database: It is responsible for storing the data that received from the previous layer and the predefined route, and then provides it to the processor.
  - Input/output device: This device allows the controller to interact with the system.
  - Processor: The processor is responsible for performing both the conflict detection and the conflict resolution algorithms. (Conflict detection algorithm is out of the scope of this paper).
  - Communication Unit: This unit is in charge of the communication process. The ground station can communicate with airplanes to perform the action using this unit.
- Application Stage: It is responsible for presenting the resolved path to the pilot via the airplane input/output device or for managing the autopilot in order to change the airplane's trajectories according to the resolved path.

At first, the system collects data via sensors. After collecting the data, the data modeling unit transfers the data into machine executable form using one of the available modeling methods, such as ontology(Baldauf et al., 2007). By performing the conflict detection algorithm, the processor decides whether there is a conflict or not. In case where there is no conflict taking place, no action will be performed by the processor and the sensors will sense new data. However, if the output of the conflict detection algorithm indicates a conflict, the processor will perform the conflict resolution algorithm (based on genetic algorithms) in order to generate the conflict-free paths. The communication unit will in turn send the resolved path to the pilot or manage the autopilot in order to change the airplane trajectories to avoid the collision.

## 5. Genetic Conflict Resolution Algorithm:

In Genetic algorithms(Goldberg, 1989, Davis, 1991, Mitchell, 1998), there are many steps to formulate the problem which are the chromosome design, fitness evaluation, crossover and mutation. The chromosome design and the fitness function (cost) will be discussed according to the type and the model that will be chosen by the user. This paper focuses on decoding two parameters for solving the problem of conflict. In 2D resolution, the system will provide a path for the conflicted airplanes using only the cross range (lateral) and down range (longitudinal) directions. Therefore, the control parameter will be the flight path angle because of the assumption that there is no need to change its height, as the heading angle is responsible for changing airplane height. In the 3D resolution, the algorithm will provide a path for the conflicted airplanes using the lateral, the longitudinal and the vertical direction. This will lead to use both the flight path angle and the heading angle. The steps of developing Genetic algorithm are:

### 5.1 Chromosome Design

The first step in developing a Genetic algorithm is the chromosome design, in which the control parameters encoded. The chromosome can handle a multiple numbers of airplanes (i.e. 1, 2, 3,..., n),according to user needs. Hence, there will be n number of subdivisions in the chromosome, each of which contains six segments used for the Delta processed by the algorithm. These segments will define the path of the airplane during 1/6 of Delta duration. In each segment, there will be a control parameter which is the differential flight path angle for 2D, and the differential flight path angle and differential heading angle for 3D. For each differential angle, the reserved number of bits is five. Figure 2 gives the overall description of our proposed chromosome for 2D and 3D resolution.

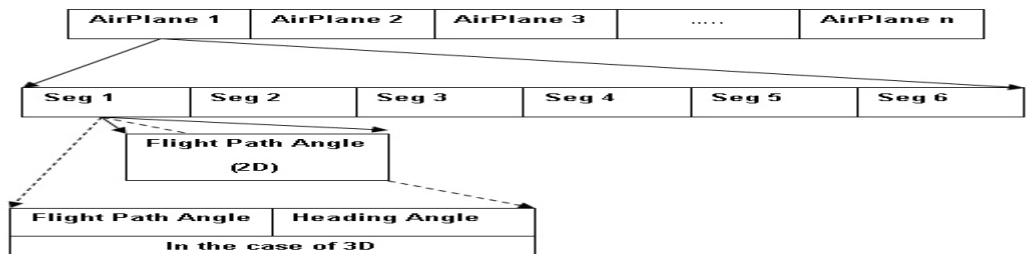


Figure 2: Chromosome Design for 2D and 3D Resolution

### 5.2 Fitness Evaluation

The second step in developing Genetic algorithm is the fitness evaluation. The algorithm resolves the conflict by searching for a flight path by maximizing the fitness function. In addition, it seeks the best resolved path

according to the user predefined preferences (i.e. minimum time). The fitness function can be expressed as follows:

$$\text{Total Fitness} = f(\text{Fit}_{\text{Conflict}}, \text{Fit}_{\text{Thrust valid}}, \text{Fit}_{\text{Load valid}}, \text{Fit}_{\text{Reaching}}, \text{Fit}_{\text{Fuel}})$$

The above fitness functions can be divided into validity of the resolved path and the fitness due to the optimality.

### 5.2.1 Conflict Fitness

The chromosome is responsible for resolving the conflict fitness, as the resulted chromosomes are always the ones that will never have a conflict in their resolved paths. Therefore, the mathematical formulation describing this fact is:

$$\text{Fit}_{\text{Conflict}} = \begin{cases} 1 & \text{if no conflict} \\ 0 & \text{otherwise} \end{cases} \quad \dots (8)$$

### 5.2.2 Fitness due to Thrust Validity

Each airplane operates with minimum and maximum thrust values, if the thrust exceeds or falls under these operating limits, there will be an error and the fitness will be zero. Otherwise the chromosome will pass for testing by the predefined fitness functions as in the following Equation:

$$\text{Fit}_{\text{Thrust valid}} = \begin{cases} 1 & \text{if (Min Thrust < Thrust < Max Thrust)} \\ 0 & \text{otherwise} \end{cases} \quad \dots (9)$$

### 5.2.3 Fitness due to Load Factor

The load factor of a given airplane in a given condition of flight is defined as the lift divided by the weight. It is denoted by  $n$ , as in the following Equation [26]:

$$n = \frac{L}{W} \quad \dots (10)$$

Referring to Equations (2) and (3),squaring and adding these two equations yield:

$$L^2 = \left( \dot{\chi} m V \cos \gamma \right)^2 + \left( m V \dot{\gamma} + g m \cos \gamma \right)^2 \quad \dots (11)$$

Finally:

$$n = \frac{\sqrt{\left( \dot{\chi} m V \cos \gamma \right)^2 + \left( m V \dot{\gamma} + g m \cos \gamma \right)^2}}{mg} \quad \dots (12)$$

Alternatively, it can be written as:

$$n = \frac{\sqrt{\left( \dot{\chi} V \cos \gamma \right)^2 + \left( V \dot{\gamma} + g \cos \gamma \right)^2}}{g} \quad (13)$$

If the load factor exceeds the predefined limits, the fitness due to the load factor will be zero, otherwise the fitness will be equal to 1.

$$\text{Fit}_{\text{load}} = \begin{cases} 1 & \text{Min Load < Load < Max Load} \\ 0 & \text{Otherwise} \end{cases} \quad \dots (14)$$

The above prevention fitness functions are of the type validity. While, the following two fitness functions are of the type optimality.

#### 5.2.4 Reaching Fitness

This section presents the fitness due to the end of the reaching for the resolved path to the end of the original path, as represented in the following Equations:

$$\text{Deviation} = \sqrt{(x_o - x_m)^2 + (y_o - y_m)^2} \quad \text{for 2D} \quad . \quad (15)$$

$$\text{Deviation} = \sqrt{(x_o - x_m)^2 + (y_o - y_m)^2 + (z_o - z_m)^2} \quad \text{for 3D} \quad \dots \quad (16)$$

Where, **o** and **m** represent the original and the modified paths respectively.

The smallest the Deviation, the maximum the fitness goes to the chromosome. The optimality index for the reaching fitness is described by the following Equation:

$$\text{Reaching Fitness} = \sum_{i=1}^n \text{Fit}_i \quad . \quad (17)$$

Where n is number of airplanes and  $\text{Fit}_i$  is

$\text{Fit}_i = \text{Max Deviation} - \text{Deviation}_i$

#### 5.2.5 Fitness due to Fuel Consumption:

The fuel consumption in the airplanes measured as the thrust applied during the flight. The minimum thrust will result in minimum fuel consumption. Hence, the thrust applied has many factors which affect its calculation, from Equations(1)-(7) we can calculate the relation between the control parameters given by the Genetic chromosome and the thrust applied by the airplane, as follows:

(Note: the velocity of the airplane is assumed constant)

$$D = q s c_{d_o} + \frac{k L^2}{q s} \quad \dots \quad (19)$$

where(Saksena et al., 1984):

q is the dynamic pressure, s is the cross sectional area, and  $c_{d_o}$  is the zero lift coefficient.

Substituting Equation (11) into Equation (13) and then into Equation (1) yields:

$$T = m g \sin \gamma + q s c_{d_o} + \frac{k}{q s} \left( \left( \dot{\chi} m V \cos \gamma \right)^2 + \left( m V \dot{\gamma} + g m \cos \gamma \right)^2 \right) \quad (20)$$

The required maneuver is in 2D. Therefore, the flight path angle will be zero, and the thrust Equation will be as follows:

$$T = q s c_{d_o} + \frac{k}{q s} \left( \left( \dot{\chi} m V \right)^2 + (g m)^2 \right) \quad \dots \quad (21)$$

The fitness due to thrust consumption will be at its maximum value whenever the thrust applied is low and vice versa. The thrust Equation in 3D will be the same as in Equation (19).

The index parameter for the fuel consumption described by the following Equation:

$$\text{Fuel Fitness} = K \left( \text{Max Thrust Sum} - \sum_{i=1}^n \sum_{t=1}^{\tau} \text{Thrust}_i \right) \quad \dots \quad (22)$$

where K is a constant, n is the number of air planes , and  $\tau$  is the number of time steps during each delta duration.

After discussing all the fitness functions, there will be one value for the overall fitness, as follows:

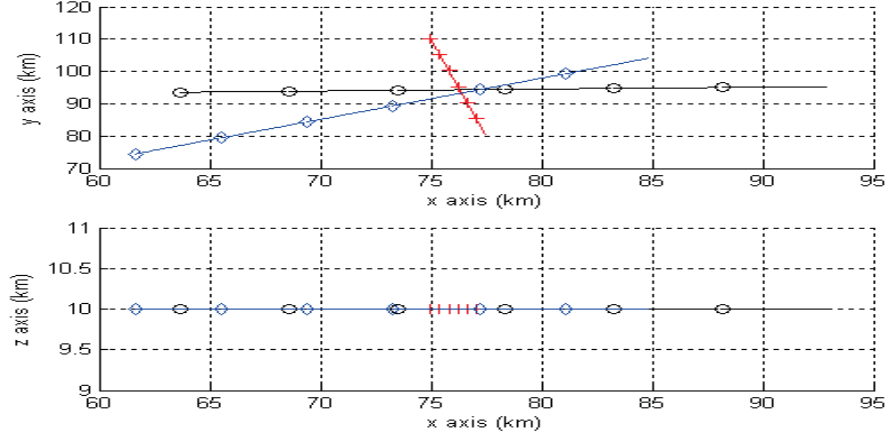
$$\text{Fit}_{\text{Total}} = \text{Fit}_{\text{Conflict}} \times \text{Fit}_{\text{Thrust Valid}} \times \text{Fit}_{\text{Load}} \sum (\text{Fit}_{\text{Reaching}} + \text{Fit}_{\text{Fuel}}) \quad \dots \quad (22)$$

### 5.3 Crossover and Mutation

The third and the fourth steps in developing a Genetic algorithm are the crossover and the mutation respectively. These steps depend on selecting the probability for crossover and mutation to be occurred. Our algorithm uses random number generated by the computer to find these probabilities.

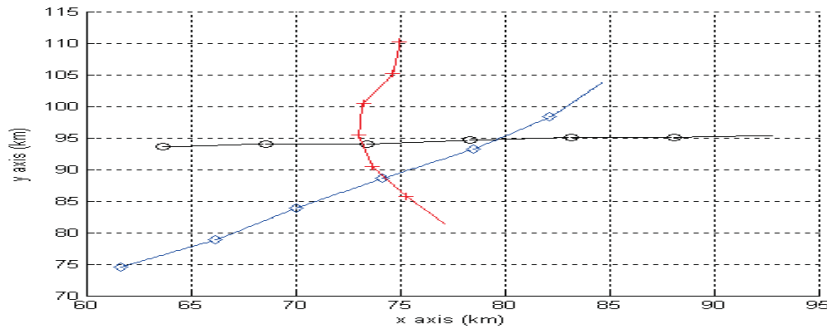
## 6. Evaluation and Results

In this section, a scenario is given to prove the validity and the effectiveness of our proposed Genetic algorithm. Figure 3 depicts the path of the three conflicted airplanes. It can be seen from the figure that each airplane's path is subdivided into six segments due to our assumption for the genetic chromosome. The airplanes assumed to be travelling in different directions at the same height. Our proposed algorithm was able to solve the conflict before it takes place by generating a new path for each airplane. As compared with the original path, the newly generated paths maintained the g-load, consumed little fuel and produced short extra distances, which in turn produced short periods of time delay.

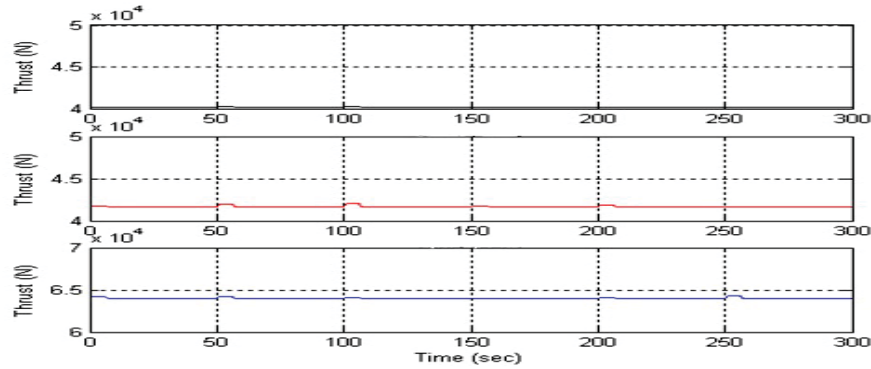


**Figure 3:** Conflict Path

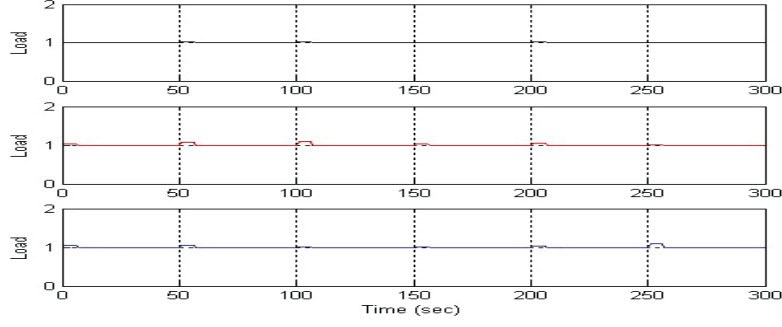
Figure 4 shows the overall resolved path using the proposed Genetic algorithm in 2D maneuvering. It can be seen from the figure that the proposed resolution of our algorithm has severely changed the path of one airplane, and it slightly changed the paths of the other airplanes. While, the thrust and the g-load remains steady in all airplanes as shown in Figure 5 and Figure 6 respectively.



**Figure 4:** Genetic Algorithm Resolved Path in 2D

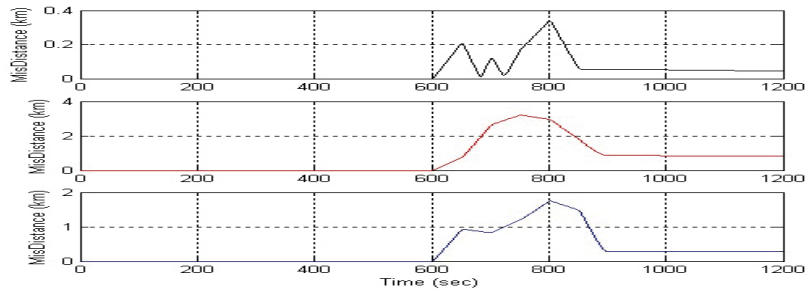


**Figure 5:** Thrust Due to Genetic Algorithm Resolution in 2D



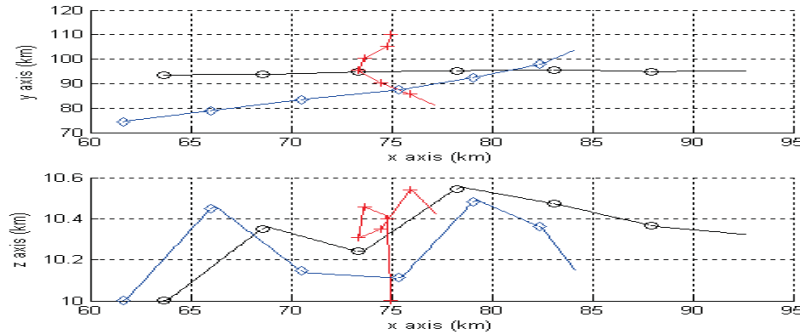
**Figure 6:** Load Due to Genetic Algorithm Resolution in 2D

However, each airplane exhibits a short period of time delay due to the change that occurred in their routes. It can be seen from Figure 7 that the extra time that has been added to the original time of airplanes is acceptable.



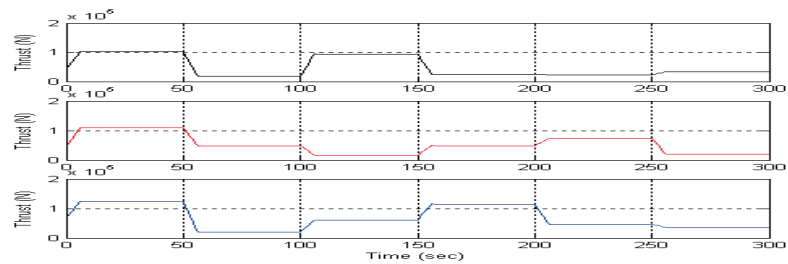
**Figure 7:** Genetic Algorithm Misdistance in 2D.

Figure 8 shows the overall resolved path using our proposed Genetic algorithm in 3D maneuvering. It can be seen from the figure that the proposed resolution of our algorithm has changed the paths of all airplanes in order to avoid the conflict. Therefore, the thrust, the g-load and the time delay have been changed.



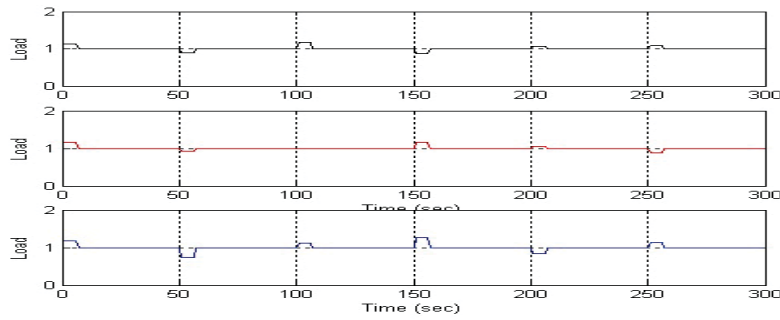
**Figure 8:** Genetic Algorithm Resolved Path in 3D

The results in Figure 9 prove the validity of our proposed algorithm in terms of maintaining an acceptable change in the thrust value. It can be seen from the figure that, there is a slight change in the thrust value for all airplanes compared to the thrust of their original paths.



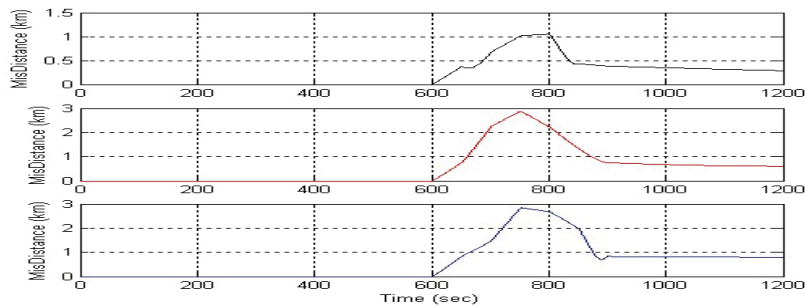
**Figure 9:** Thrust Due to Genetic Algorithm Resolution in 3D

Figure 10 depicts the change in the g-load that takes place due to the change in paths. It can be seen from the figure that, the g-load is almost acceptable from both pilot and travellers in all airplanes.



**Figure 10:** Load Due to Genetic Algorithm Resolution in 3D

In Figure 11, an illustration to the change in distance due to the change in paths is presented. It can be seen from the figure that, there are very short extra distances added to each path which in turns caused a short time delay.



**Figure 11:** Genetic Algorithm Misdistance in 3D

As shown in the results, our proposed algorithm was able to solve the conflict problem performing both 2D and 3D maneuvering. The generated path in both scenarios (2D and 3D) has met the criteria we have identified earlier, which are resolving the conflict without adding extra distance (time delay), consuming the fuel as less as possible and maintaining the g-load during changing the airplane's trajectories.

## 7. Conclusion

Conflict resolution has been recognized as a challenging problem facing air traffic management systems' controllers, especially in the dense areas. In this paper, a context-aware air traffic management system has been introduced to solve the problem of conflict by performing both 2D and 3D maneuvering. A novel context-aware architecture for the air traffic management system's ground station has been proposed. In addition, a novel conflict resolution based on Genetic algorithms has been introduced in the reasoning stage of the architecture. The results have proved the efficiency and the effectiveness of our proposed algorithm in solving the conflict by generating conflict-free paths which satisfy the requirements of fuel consumption, time delay and g-load. Our future work in this area will investigate the use of our proposed algorithm in another flight phase (i.e departure).

## References

- Baldauf, M., Dustdar, S. & Rosenberg, F. 2007. A Survey On Context-Aware Systems. *International Journal Of Ad Hoc And Ubiquitous Computing*, 2, 263-277.
- Blakelock, J. H. 1991. *Automatic Control Of Aircraft And Missiles*, Wiley New York.
- Chaloulos, G., Hokayem, P. & Lygeros, J. Distributed Hierarchical Mpc For Conflict Resolution In Air Traffic Control. *American Control Conference (Acc)*, 2010, 2010. Ieee, 3945-3950.
- Cobano, J. A., Alejo, D., Ollero, A. & Viguria, A. Efficient Conflict Resolution Method In Air Traffic Management Based On The Speed Assignment. *Proceedings Of The 2nd International Conference On Application And Theory Of Automation In Command And Control Systems*, 2012. Irit Press, 54-61.
- Davis, L. 1991. *Handbook Of Genetic Algorithms*.
- Dey, A. K., Abowd, G. D. & Salber, D. 2001. A Conceptual Framework And A Toolkit For Supporting The Rapid Prototyping Of Context-Aware Applications. *Human-Computer Interaction*, 16, 97-166.

- Emami, H. & Derakhshan, F. An Overview On Conflict Detection And Resolution Methods In Air Traffic Management Using Multi Agent Systems. *Artificial Intelligence And Signal Processing (Aisp)*, 2012 16th Csi International Symposium On, 2012. Ieee, 293-298.
- Gao, Y., Zhang, X. & Guan, X. Cooperative Multi-Aircraft Conflict Resolution Based On Co-Evolution. *Instrumentation & Measurement, Sensor Network And Automation (Imsna)*, 2012 International Symposium On, 2012. Ieee, 310-313.
- Goldberg, D. E. 1989. *Genetic Algorithms In Search, Optimization, And Machine Learning*, Addison-Wesley Reading Menlo Park.
- Indulska, J. & Sutton, P. Location Management In Pervasive Systems. *Proceedings Of The Australasian Information Security Workshop Conference On Acsw Frontiers 2003-Volume 21*, 2003. Australian Computer Society, Inc., 143-151.
- Karr, D. A., Vivona, R., Roscoe, D., Depascale, S. M. & Consiglio, M. Experimental Performance Of A Genetic Algorithm For Airborne Strategic Conflict Resolution. *Proceedings Of The Eight Usa/Europe Air Traffic Management Research And Development Seminar (Atm2009)*, Napa (California, Usa), 2009.
- Khedkar, A. & Kumar, V. Self Synchronization Based Air Traffic Control And Collision Avoidance System. *Digital Avionics Systems Conference (Dasc)*, 2011 Ieee/Aiaa 30th, 2011. Ieee, 1d4-1-1d4-8.
- Kuchar, J. K. & Yang, L. C. 2000. A Review Of Conflict Detection And Resolution Modeling Methods. *Intelligent Transportation Systems, Ieee Transactions On*, 1, 179-189.
- Malaek, S., Alaeddini, A. & Gerren, D. 2011. Optimal Maneuvers For Aircraft Conflict Resolution Based On Efficient Genetic Webs. *Aerospace And Electronic Systems, Ieee Transactions On*, 47, 2457-2472.
- Menon, P., Sweriduk, G. & Sridhar, B. 1999. Optimal Strategies For Free-Flight Air Traffic Conflict Resolution. *Journal Of Guidance, Control, And Dynamics*, 22, 202-211.
- Mitchell, M. 1998. *An Introduction To Genetic Algorithms*, Mit Press.
- Peyronne, C., Delahaye, D., Lapasset, L. & Mongeau, M. Air Traffic Conflict Resolution Via B-Splines And Genetic Algorithms. *Global Optimization Workshop*, 2010. 127-130.
- Saksena, V., O'reilly, J. & Kokotovic, P. V. 1984. Singular Perturbations And Time-Scale Methods In Control Theory: Survey 1976–1983. *Automatica*, 20, 273-293.
- Vela, A., Feigh, K. M., Solak, S., Singhose, W. & Clarke, J. 2012. Formulation Of Reduced-Taskload Optimization Models For Conflict Resolution. *Systems, Man And Cybernetics, Part A: Systems And Humans, Ieee Transactions On*, 42, 1552-1561.
- Vela, A. E., Solak, S., Feron, E., Feigh, K., Singhose, W. & Clarke, J.-P. A Fuel Optimal And Reduced Controller Workload Optimization Model For Conflict Resolution. *Digital Avionics Systems Conference*, 2009. Dasc'09. Ieee/Aiaa 28th, 2009. Ieee, 3. C. 3-1-3. C. 3-16.
- Zhang, J., Wu, J. & Song, T. Study Of Multi-Aircraft Conflict Resolution And Algorithm Optimization Based On Genetic Algorithm. *2014 International Conference On Computer, Communications And Information Technology (Ccit 2014)*, 2014. Atlantis Press.

# Post Implementation Of ERP System

Imran Batada<sup>1</sup> and Asmita Rahman<sup>2</sup>

<sup>1</sup>Institute of Business Administration, Karachi, Pakistan

<sup>2</sup>University Of Georgia, Athens USA

[imran.batada@khi.iba.edu.pk](mailto:imran.batada@khi.iba.edu.pk)

[asmitab@uga.edu](mailto:asmitab@uga.edu)

**Abstract:** The trend of ERP Implementation is rapidly increasing not only in the developed countries but also in the developing countries. However, sometimes a system that works well in one country and industry may not work well in other countries. This research identifies the problems faced today in the post implementation phase of ERP in Pakistan. This research is a continuation work of a previous research published in the paper 'Measuring System Performance & User Satisfaction after Implementation of ERP' (Batada, I., Rahman, A., (2012)). This research further advances the previous work by focusing on the implementation of five (5) public sector universities who have implemented PeopleSoft ERP System within the last four (4) years. A custom questionnaire criterion was developed and floated in all those universities. This research uses 'Balance Score Card' approach to evaluate the performance of ERP System. The outcome of the survey clearly indicates that the ERP System was not able achieve the desired results. This study will help the universities to identify the problem areas resulting in increased overall utilization. The results obtained in this research can also benefit other universities in the region.

**Keywords:** ERP System Performance Measurement, ERP performance evaluation, ERP User Satisfaction, Balance Scorecard

---

## 1. Introduction

Enterprise Resource Planning (ERP) System can be defined as a business management software that helps organization(s) improve the business process by following the best practices. The adoption of business process reengineering is tough to digest at the beginning of roll out; however, it has long term affective impacts. ERP System maintains all the data of the organization in a central database system. This helps organization in not only fetching results in real time but also in taking timely decisions. Successful implementation of ERP can only be measured by its utilization factor. Low utilization does not always mean that selection of the product is not good, since the same product may work well in other organizations. The utilization measurement exercise identifies the areas that are working well and the areas that need improvement.

## 2. Related Works

ERP systems are complex and large scale systems developed to integrate business logic and processes. Their implementations are expensive in terms of both; time and money. Utilization of an ERP system is a critical factor when evaluating the success of the implementation. Several researches have been done to accomplish a good measure for its performance and implementation success.

Althonayan, M., Papazafeiropoulou, A., (2011) used a measure designed as a combination three information system models: Delone and McLean's IS success, Task Technology Fit (TTF) by Goodhue and End User Computing Satisfaction (EUCS) by Doll and Torkzadeh. T.

Another approach was taken by Ho, C., (2007) to measure the performance on ERP in supply chain area. He used an integrated method, total related cost measurement, to evaluate the performance of a three-echelon ERP based supply chain system.

## 3. Relevant Methods for ERP Performance Measurement

ERP System's effectiveness is complex to measure, since it covers the entire organization of Information System. There are a number of researches that have already been conducted to evaluate different methods for measuring the overall performance of ERP System.

One of the most cited information model is the DeLone McLean I/S Success Model. They have focused on Quality of Information, System Quality, Quality of Service, Intention to use, User satisfaction and net benefits dimensions in their model. This method defines all the 6 categories relationship and their interdependencies.



Another famous model is the Markus & Tanis (2000), they have focused from the project planning stage to the post implementation results. According to them, each ERP Implementation is unique in nature and experience of different organization may vary.

Another widely used method is the Balanced Scorecard Method. This method focuses on the four major areas of the organization. We have also used this method for evaluating the performance of 5 universities. This method covers the following four perspectives; financial/cost perspective, customer perspective, internal processes perspective and innovation and learning perspective. This method was chosen because it considers several areas that impact any organization and enables to evaluate based on each and every area.

#### **4. Selection of Balanced Scorecard for ERP Performance System**

The Balance scorecard method was chosen and was carried out in all the five selected universities of Pakistan. All the selected universities have recently implemented ERP system. This evaluation helped us identify the strengths and weaknesses of the implementations. It also provides us a way to access the overall advantage of the ERP system implementation

#### **5. Survey Design**

In order to evaluate the performance of the ERP system, a questionnaire survey was developed. The survey consisted of 48 questions. These questions comprised of all the various aspects of the ERP system to better evaluate each and every area. An online survey consisting of the questionnaire was distributed among the five universities. The participants of the survey included all major stakeholders and results were kept anonymous. The questions of the survey were distributed among the four major perspectives i.e. Internal Business perspective, Learning and Growth perspective, and financial perspective and Customer perspective.

#### **6. Research Method- Balanced Scorecard**

A balanced scorecard method is a projection and management system used by various organizations in order to monitor and evaluate performance. It is a full strategic planning system enabling organizations to evaluate their performance against global goals.

Balanced scorecard method evaluates on the basis of four major perspectives. The following four perspectives were considered Financial, Internal Business, Customer & Learning and Growth. Based on each perspective, data was gathered and analysed. The following diagram shows the balanced scorecard method:

#### **7. Balanced Scorecard Implementation**

All survey results were recorded for data analysis. The survey consisted a total of 48 questions. All these questions were divided among all the perspectives. The same questionnaire of our earlier research paper "Measuring System Performance & User Satisfaction after Implementation of ERP" has been used. All the major stakeholders participated in the survey. The questions were divided as follows among the perspectives:

Learning and Growth Perspective: 12 questions

Internal Business and Growth Perspective: 16 questions

Customer Perspective: 12 questions

Financial Perspective: 8 questions

#### **8. Performance Evaluation**

##### **8.1 Perspective # 1: Learning and Growth**

This perspective consists of three major matrices; for better classification of the questions. These three matrices include the questions below:

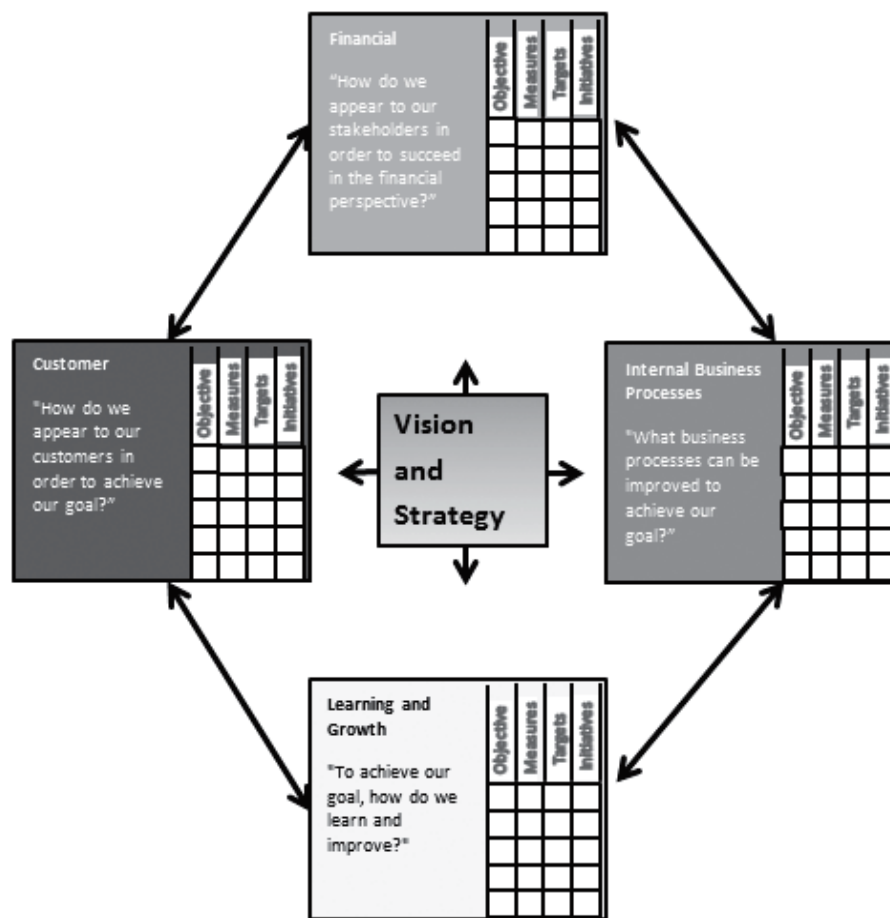
###### **8.1.1 Matrix 1**

Question No 1: Is the ERP Data accurate?

Question No 4: Are you comfortable to rely on the system information?

Question No 3: Are you or your department capable to take care of your reporting requirements?

Question No 1: Did the ERP System increase your productivity?



**Figure 1:** Balance Scorecard approach for performance measurement (Robert, Kaplan, David & Norton 1996)

**Table 1:** Matrix 1 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	4.25	4.5	4.25	3.75	4.25
Score of Matrix 1:	0.85	0.9	0.85	0.75	0.85

### 8.1.2 Matrix 2

Question No 5: Are you getting up-to-date information from the System?

Question No 6: Are you getting required information instantly from the System?

Question No 11: The information which could only be obtained only from other departments earlier, is now available via the system

Question No 2: Strong Information Systems (IS) Department

**Table 2:** Matrix 2 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	3.75	4.75	3.25	4.5	3.5
Score of Matrix 2:	0.75	0.95	0.65	0.9	0.7

### 8.1.3 Matrix 3

Question No 3: Technical knowhow of the ERP team

Question No 5: Strong and meaningful training programs

Question No 11: A good understanding of the concept of ERP

Question No 12: A good IT infrastructure in place already

**Table 3:** Matrix 3 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	3.75	4.75	3.25	4.5	3.5
Score of Matrix 3:	0.75	0.95	0.65	0.9	0.7

### 8.1.4 Performance:

**Table 4:** Performance of the first perspective

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Performance (Learning and growth Perspective)	0.7992	0.949	0.782	0.782	0.782

## 8.2 Perspective # 2: Internal Business Perspective:

This perspective consists of four major matrices; for better classification of the questions. These three matrices include the questions below:

### 8.2.1 Matrix 1:

Question No 2: Is the System efficient to take your needs?

Question No 1: Is the ERP System advantageous?

Question No 2: Does the ERP system aid your decision making process?

Question No 1: Are you getting good ERP Technical Support?

**Table 5:** Matrix 1 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	3.5	4	4	3.75	3.5
Score of Matrix 1:	0.7	0.8	0.8	0.75	0.7

### 8.2.2 Matrix 2:

Question No 2: Are you or your department being trained on time?

Question No 1: Does the ERP System ease your work load?

Question No 2: Does the ERP system provide you accurate information for your daily routine tasks?

Question No 3: Does the ERP System provide you the information for easy decision making?

**Table 6:** Matrix 2 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	3.25	4	4.25	3.75	3.75
Score of Matrix 2:	0.65	0.8	0.85	0.75	0.75

### 8.2.3 Matrix 3

Question No 10: Paper work has been reduced in the organization  
of ownership improved among the functional heads and end users  
Question No 13: The interdepartmental relationships improved  
Question No 14: Overall group / departmental productivity improved

**Table 7:** Matrix 3 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	3.75	4.5	3.5	3.25	4.25
Score of Matrix 3:	0.75	0.9	0.7	0.65	0.85

### 8.2.4 Matrix 4

Question No 4: Better coordination in between different departments  
Question No 6: Streamlined the business processes  
Question No 7: Increased overall satisfaction with business processes  
Question No 8 Improve the organizational processes as a whole

**Table 8:** Matrix 4 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	4	5	4.5	1.25	4
Score of Matrix 4:	0.8	1	0.9	0.25	0.8

### 8.2.5 Performance:

**Table 9:** Performance of the second perspective

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Performance	0.725	0.875	0.8125	0.6	0.775

## 8.3 Perspective # 3: Customer Perspective

This perspective consists of three major matrices; for better classification of the questions. These three matrices include the questions below:

### 8.3.1 Matrix 1:

Question No 3: Did you get any benefit out of resource utilization?  
Question No 3: Are you getting timely reports from the System?  
Question No 10: User's involvement and participation  
Question No 3: For how long do you use the system daily

**Table 10:** Matrix 1 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	4	4.5	4	3.75	3.75
Score of Matrix 1:	0.8	0.9	0.8	0.75	0.75

### 8.3.2 Matrix 2:

Question No 4: Are you or your team capable to handle the ERP Functional Side?

Question No 4: Is the ERP System user friendly?

Question No 7: What is your opinion about the overall performance of the System?

Question No 9: ERP simplifies user task

**Table 11:** Matrix 2 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	3.75	4.75	3.25	3.75	4
Score of Matrix 2:	0.75	0.95	0.65	0.75	0.8

#### 8.3.3 Matrix 3:

Question No 1: Top management support

Question No 4: Functional areas support

Question No 7 Overall support for the ERP in the organization

Question No 9: Your opinion about ERP software selected

**Table 12:** Matrix 4 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	4.24	5	3	4.25	3.5
Score of Matrix 3:	0.848	1	0.6	0.85	0.7

#### 8.3.4 Performance

**Table 13:** Performance of the third perspective

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Performance	0.799	0.949	0.682	0.782	0.749

### 8.4 Perspective # 4: Financial Perspective

This perspective consists of two major matrices; for better classification of the questions. These three matrices include the questions below:

#### 8.4.1 Matrix 1:

Question No 8 ERP helps in saving end user time

Question No 2: Was there any staff reduction?

Question No 1: Was there a decrease in overall operating cost after implementing the System?

Question No 8: Efficient change management

**Table 14:** Matrix 1 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	3.25	4.5	3.25	3.5	4
Score of Matrix 1:	0.65	0.9	0.65	0.7	0.8

#### 8.4.2 Matrix 2:

Question No 6: Strategic planning for the ERP Project

Question No 5: Forecasting process has been improved

Question No 2: Are you dependent on the ERP System for your daily operational activities?

Question No 3: Has the system improved your performance?

**Table 15:** Matrix 2 scores

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Avg. Weightage of the questions:	5	5	5	5	5
Avg. response score for this matrix:	4	5	3.75	2.75	4
Score of Matrix 1:	0.8	1	0.75	0.55	0.8

#### 8.4.3 Performance

**Table 16:** Performance of the fourth perspective

	Univ. 1	Univ. 2	Univ. 3	Univ. 4	Univ. 5
Performance	0.727	0.952	0.7	0.625	0.8

### 8.5 Performance Chart

**Table 17:** Overall performance

	University 1	University 2	University 3	University 4	University 5
Learning and Growth Perspective	0.7992	0.949	0.782	0.782	0.782
Internal Business Perspective	0.725	0.875	0.8125	0.6	0.775
Customer Perspective	0.799	0.949	0.682	0.782	0.749
Financial Perspective	0.727	0.952	0.7	0.625	0.8
Overall Avg. Score	0.76255	0.93125	0.744125	0.69725	0.7765

### 8.6 Evaluation Criteria

Here is scale, which was used to measure and analyse the performance of our system:

**Table 18:** Scale

Scale:		
	Excellent	0.86 to 0.10
	Good	0.66 to 0.85
	Fair	0.51 to 0.65
	Poor	0.0 to 0.5

## 9. Critical Analysis

As we can see from the above scores, four out of five universities scored under 0.80 on a scale of 0-1. Based on the scale shown above, the overall performance score of four out of five universities fell under the ranking of “Good”. This clearly indicates that about 80% of the universities in the considered set, need considerable improvement in order to achieve the desirable ‘excellent’ score based on the scale above. The perspective that scored the least was the Internal Business Perspective with an overall average of 0.757. Based on this evaluation, it can be concluded that this perspective requires the most improvement. This perspective can be improved by streamlining the business processes. Based on the results obtained in the survey, one of the weak areas that need improvement is providing proper training to the users.

The perspective that scored second last was the financial perspective with an overall average of 0.7608. This perspective is a clear indication of the higher cost of implementation of the ERP system. This perspective can

be improved by reducing the over-all cost and expenses of the implementation and by providing greater value to system's users by decreasing the routine operational costs.

Customers Perspective showed 0.792 score, which is also good rating but can be improved. This perspective can be improved by providing support to users in the functional area, by conducting inter departmental meetings etc. As indicated, ERP System is a centralized system that is connected with all the departments therefore; the interdepartmental support is essential to help improve the ranking of this phase.

Learning and growth perspective scored the highest. However, the average score of this perspective was still in the same category of Good with the score of 0.818.

## **10. Conclusion**

Measurement of ERP System's performance is essential for any organization that has implemented an ERP system. Such an analysis, not only helps the organization to evaluate the success of the implementation, but also gives a high level idea on specific areas of improvement. This helps the organization identify their strengths and highlight their weaknesses. It is necessary for all the stakeholders to participate in the evaluation process to get accurate results.

ERP System incurs a lot of cost and effort; therefore, it is necessary to give proper consideration to the utilization factor. Underutilized product results in the overall failure of the project and will not benefit the organization. It is recommended to conduct several surveys every year in order to get proper measurement of utilization.

## **11. Future Works**

Areas of future research may include conducting a similar survey every year and compiling the results. This will not only give a long term trend of utilization but will also consider the fact the performance may be affected more in the first couple years of rollout.

Another further area of research may include changing the weightage of each perspective depending on the business model of a particular organization.

## **References**

- Batada, I., Rahman, A., (2011). Selection, Implementation & Post Production of an ERP System. Proceedings of the 2nd International Conference on Information Management and Evaluation.
- Batada, I., Rahman, A., (2012). Measuring System Performance & User Satisfaction after Implementation of ERP. Proceedings of Informing Science & IT Education Conference (InSITE) 2012.
- Chen, S., Lin, Y., (2008). An Evaluation Method For Enterprise Resource Planning Systems. Journal of the Operations Research Society of Japan
- Dahne, S., Hohne, J., & Tangermann M. (2011). Adaptive Classification Improves Control Performance in. ERP-Based BCIs. Proceedings of the 5th International BCI Conference 2011.
- Ho C. (2007). Measuring system performance of an ERP-based supply chain. International Journal of Production Research
- Hyvonen J.(2007) Strategy, performance measurement techniques and information technology of the firm and their links to organizational performance. Management Accounting Research 18
- Robert, S., Kaplan., David P., Norton, (1996 ) Using the Balanced Scorecard as a Strategic Management System, Harvard Business Review.
- Rosemann, M., Wiese, J., (1999), Measuring the Performance of ERP Software– a Balanced Scorecard Approach. 10th Australasian Conference on Information Systems, 1999. .
- Wei, C., Liou, T., Lee, K. (2008). An ERP performance measurement framework using a fuzzy integral approach. Journal of Manufacturing Technology Management.

# An Analysis of Canadian Armed Forces Historical Release Rates by Location

Din Begovic

Director General Military Personnel Research and Analysis, Ottawa, Canada

[din.begovic@forces.gc.ca](mailto:din.begovic@forces.gc.ca)

**Abstract:** The Department of National Defence (DND) conducts operational research to monitor the health and long-term sustainability of military occupations through changes in the yearly personnel release statistics. Military occupations whose yearly release numbers consistently increase over several years are analysed in detail so that corrective action can be taken, if required. A 2013 study found that the Canadian Armed Forces (CAF) attrition rates in Alberta had increased between fiscal years (FY) 11/12 and 12/13. Anecdotally, occupation managers have hypothesized that Alberta's strong economy may be causing higher than expected numbers of CAF personnel to release from the military to seek employment in the private sector. This paper describes an analytic approach used to investigate whether releases in Alberta have increased during the recent past, and, if so, to identify the location and characteristics of the releases. Results of the preliminary analysis presented in this paper suggest that release rates have indeed increased in Alberta in the last two years. However, changes in yearly release rates in Alberta appear to be occupation and location dependent. As an example, combat engineers, one of the military occupations in Alberta, appear to have been releasing at a higher rates. Combat engineers in Alberta also appear to be more likely to release earlier in their careers than combat engineers in other provinces. Findings from this and additional studies could be used for the future management of CAF military occupations.

**Keywords:** Canadian Armed Forces, Military Occupation Management, Release Rate Analysis

---

## 1. Background

Director General Military Personnel Research and Analysis (DGMPPRA) conducts operational research in the personnel domain. One research method for monitoring the health and long-term sustainability of CAF occupations involves assessing yearly release rates. Examining yearly release rates allows for better management of military occupations with respect to recruitment, training, and promotion of members. This research is especially important for managing those occupations that are currently under-staffed or that have very long training times for new recruits. It has recently been observed that CAF release rates in Alberta have increased significantly over historical rates. One potential explanation is that the strong Alberta economy has induced CAF members to leave the military for private sector employment. Anecdotally, the oil industry in Edmonton has been cited as a particularly likely draw for CAF members who possess skills that can be transferred to that industry. Without a formal study, however, it has been difficult to determine whether the observed increases in release rates truly exist, and if so, to what extent.

In July 2013, DGMPPRA's Workforce Modelling and Analysis Team conducted a preliminary study confirming that the release rates in Alberta had increased significantly between FYs 11/12 and 12/13. (Latchman, 2013) The study presented in this paper compliments the July 2013 study by updating the study and investigating the characteristics of historical releases of non-commissioned members (NCMs). Specifically, this paper examines the exact locations (cities/bases) and military occupations from which the releases have occurred. This additional level of detail will add to the evidence required to determine whether the increased release rates in Alberta are occurring across the province and for all NCM occupations in the province, or only from specific locations and occupations. Finally, this study serves as a preliminary investigation into potential demographic explanations for the increased release rates in Alberta.

## 2. Methodology

Data on CAF personnel (demographics and releases) are stored in the Human Resource Management System (HRMS) database. The data for this study are based on HRMS extracts, imported into a Microsoft Access™ database by DGMPPRA's technical support team and then imported into Microsoft Excel™ for further analysis by the author. In this study, the term "releases" denotes those CAF members who have left the Regular Force component of the CAF for either voluntary or non-voluntary reasons. Only data from FY 04/05 to FY 13/14 were used in this study because of deficiencies in the location field in the older data in the MS Access™ database. Additionally, this study does not differentiate between trained and untrained (recruits) members. Finally, this paper only presents the NCM CAF population analysis due to space constraints and because the



NCM release rates in Western Canada have been more noticeably higher than in the rest of Canada. (Fang and Bender, 2009) (Latchman, 2013)

Next, the data in Microsoft Excel™ were filtered into logical groupings to better structure the analysis. Release data were first grouped by province of release. (Note that Saskatchewan, Prince Edward Island, Newfoundland, and the Canadian territories have not been analysed because they experienced extremely small numbers of yearly CAF releases due to their small population.) This grouping was then subdivided into the location of release. Several individual locations listed in the dataset were collated because of their geographical proximity to one another. Cold Lake and Medley (Alberta), for example, were grouped together. Next, the data were organized by military occupation to provide more detail about which military occupations in Alberta had experienced the most pronounced changes in historical release rates. Finally, data on historical years of service (YOS) and age profiles were selected in order to provide additional insight into the demographics of the members that have released. Note that YOS is an important demographic in release studies because releases have been found to be directly related to YOS (Otis and Straver, 2008). For example, many CAF members chose to leave the military at 20 YOS because they were eligible to receive an immediate annuity at that time.

Two types of analysis were conducted on the prepared data in Microsoft Excel™. First, historical release volumes by province were tabulated for every year beginning with FY 04/05 to provide an assessment of the yearly movement of NCMs out of the CAF. However, a direct comparison of provincial release volumes does not account for the differing provincial populations. In order to control for population size differences so that comparisons between provinces can be made, release volumes were converted to release rates. As with previous release rate studies, yearly historical release rates were calculated by dividing the number of releases in one year by the total population at the beginning of the year. (Okazawa, 2007) Release rates were calculated for every FY from 04/05 to 13/14. Because this study was conducted before the end of FY 13/14, release rates for FY 13/14 were estimated based on actual release data from April 2013 to January 2014 and forecasted for February and March. Consequently, all release volumes and rates for FY 13/14 presented in this study are subject to change.

### **3. Results**

Table 1 presents the yearly release volumes by province. The figures presented in this table are not only useful for observing general changes in release numbers over time, but also indicate the general movement of NCM personnel out of the CAF. The provinces are listed west to east rather than alphabetically. This and other tables on release rates and volumes also provide the ten-year average population (a.p.) for each province in order to bring attention to the size differences between the provinces. The same approach was used for presenting the population sizes of the various locations and military occupations. Further, years in which the release volumes were more than 10% higher than the previous year have been shaded in Table 1 to assist with the identification of increasing trends. (Note, however, that the 10% cut-off percentage was chosen arbitrarily to help identify the most pronounced year-to-year changes.) Findings presented in Table 1 show that all the provinces analysed in this study had at least one year where the release volume increased by 10% or more over the previous year. While several provinces had years where release volume changed significantly from one year to the next, only Alberta and Nova Scotia showed large increases in release volumes in the last two years. It is also important to notice the numerous instances of year-over-year decreases in release volumes. In light of these fluctuations, it is difficult to determine whether the significant increases in release volumes for FYs 12/13 and 13/14 in Alberta will continue. Finally, FY 13/14 release volumes for all the provinces except British Columbia and Manitoba are forecast to exceed FY 12/13 release volumes. Note that FY 12/13 population sizes of all the provinces are within 3% of the FY 13/14 sizes. This finding eliminates increased population sizes as a contributing factor for increased release volumes in these last two years; it does not, however, account for changes to population sizes between FY 04/05 and FY 12/13.

**Table 1:** Historical NCM release volumes by province and fiscal year.

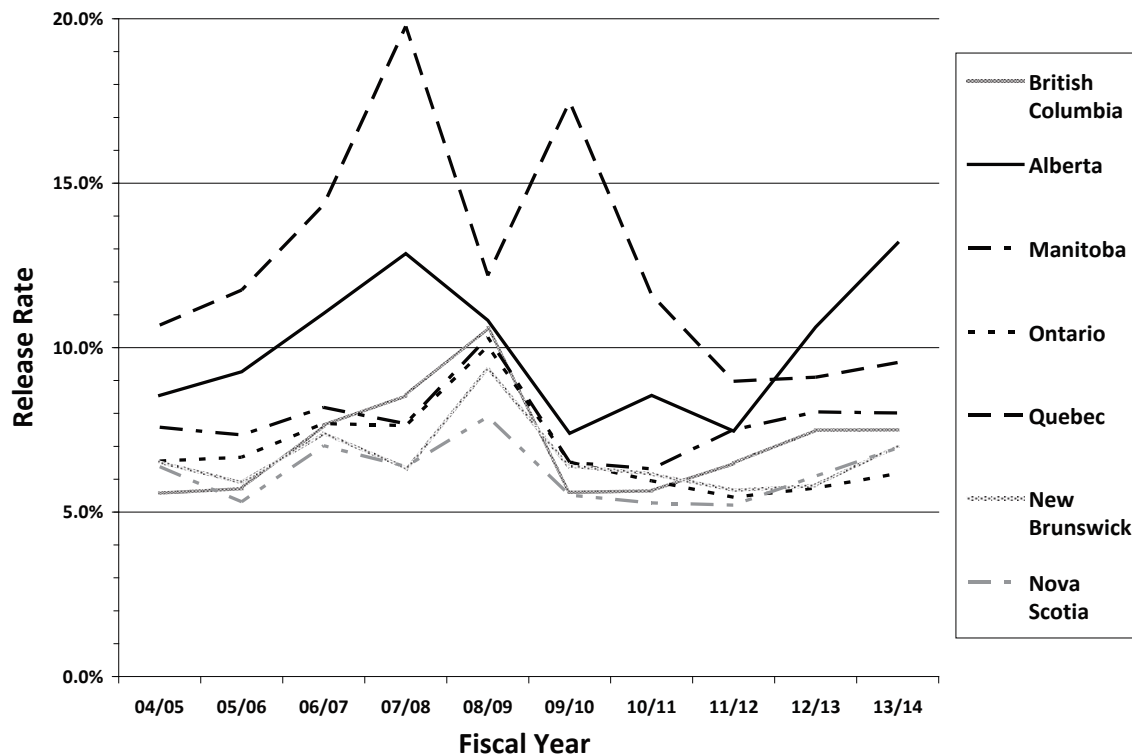
FY	British Columbia	Alberta	Manitoba	Ontario	Quebec	New Brunswick	Nova Scotia
	a.p. = 4,399	a.p. = 6,398	a.p. = 2,184	a.p. = 15,972	a.p. = 8,896	a.p. = 3,654	a.p. = 6,717
04/05	262	561	165	951	839	210	441
05/06	264	590	162	958	941	193	370
06/07	342	699	176	1,177	1,215	254	485
07/08	366	804	169	1,206	1,771	217	427
08/09	430	673	218	1,622	1,134	334	511
09/10	224	479	137	1,079	1,676	232	351
10/11	236	569	138	1,016	1,114	259	345
11/12	288	498	168	933	812	226	349
12/13	342	693	181	958	829	230	413
13/14	334	825	181	1,025	909	283	470

Table 2 provides a different perspective on the historical CAF releases because it shows release rates as opposed to release volumes. Note that the years where the release rates increased for two or more consecutive years (regardless of size of increase) are shaded; the number of consecutive years was chosen arbitrarily to make the most pronounced increasing trends more visible. Specifically, with respect to the last two years, observe that Alberta, Ontario, Quebec, New Brunswick and Nova Scotia have all experienced an increase in yearly release rates in the last two years. However, this increase is more pronounced in Alberta where changes in the both release rates and volumes are larger than those of other provinces. Finally, as was the case with release volumes, large fluctuations can be observed between the yearly release rates. These fluctuations increase the complexity of managing the occupations. Therefore, it is important to continuously monitor the yearly release rates for changes.

**Table 2:** Historical NCM release rates by province and fiscal year.

Fiscal Year	British Columbia	Alberta	Manitoba	Ontario	Quebec	New Brunswick	Nova Scotia
	a.p. = 4,399	a.p. = 6,398	a.p. = 2,184	a.p. = 15,972	a.p. = 8,896	a.p. = 3,654	a.p. = 6,717
04/05	5.6%	8.5%	7.6%	6.6%	10.7%	6.5%	6.4%
05/06	5.7%	9.3%	7.3%	6.7%	11.7%	5.9%	5.3%
06/07	7.6%	11.0%	8.2%	7.7%	14.3%	7.4%	7.0%
07/08	8.5%	12.9%	7.7%	7.6%	19.8%	6.3%	6.4%
08/09	10.6%	10.8%	10.3%	10.0%	12.2%	9.3%	7.9%
09/10	5.6%	7.4%	6.5%	6.5%	17.5%	6.4%	5.5%
10/11	5.6%	8.6%	6.3%	6.0%	11.6%	6.2%	5.3%
11/12	6.5%	7.5%	7.5%	5.5%	9.0%	5.7%	5.2%
12/13	7.5%	10.6%	8.0%	5.7%	9.1%	5.8%	6.1%
13/14	7.5%	13.2%	8.0%	6.2%	9.5%	7.0%	7.0%

Figure 1 presents the content of Table 2 in a diagrammatical form, which makes for easier comparisons of provincial release rates. Observe that Alberta's release rates are higher than those of almost all of the other provinces analysed in this study. Other than in FYs 12/13 and 13/14, Quebec's release rates have been significantly higher than the release rates of the other provinces, including Alberta. Release rates in Quebec are greatly influenced by the number of recruits who do not successfully complete the Basic Military Qualification (BMQ) course, which is taught in St. Jean, Quebec. (Fang and Bender, 2009) Finally, release rates for all of the provinces showed increases and decreases over the past ten years, which entails, once again, that the increase in the past three years could eventually reverse direction. It is important to note, however, that release rates are affected through numerous internal and external factors that are not directly discussed in this paper. For example, release rates that are deemed too high can be decreased through the implementation of CAF retention policies.



**Figure 1:** Comparison of historical NCM release rates by province.

This study also examined whether the increased release rates in Alberta were exclusive to one or more locations. Table 3 highlights the most significant changes in release rates observed in Alberta from FY 04/05 to FY 13/14. Note that Lethbridge, Medicine Hat, and Red Deer were not included in the analysis because these three locations had a negligible combined total of 21 releases in the past ten years. As with Table 2, release rates that increased for two or more consecutive years are shaded. Cold Lake (and Medley) is the only location in Alberta that has experienced a consistent yearly increase in release rates over the last three years, rising from 5.5% in FY 10/11 to approximately 13% in FY 13/14.

One peculiar finding is the extremely high release rate in Calgary in FY 08/09. A closer examination of Calgary's historical releases (not shown) did not immediately reveal a reason for such a dramatic increase in its release rate, given that the releases occurred across numerous occupations. However, Calgary has one of the smallest CAF populations in Alberta; therefore, small changes in yearly release volumes could have resulted in larger release rate changes and widely fluctuating release rates. This finding could also explain the 26% release rate in FY 06/07 and the 18% release rate in FYs 07/08 and 13/14. Similar fluctuations in yearly release rates also occurred in Suffield and Wainwright, both of which have small populations.

Finally, Edmonton release rates did increase significantly from FY 09/10 to FY 13/14. However, the preliminary release rate for FY 13/14 appears similar to that of FY 12/13, suggesting that the release rate might have started to plateau, which can only be determined continuing to monitor Edmonton release rates. Furthermore, out of the five Alberta locations examined, Edmonton's release rates in FY 13/14 were actually the second lowest and similar to that of Cold Lake (and Medley). Nonetheless, Edmonton's large population in Alberta results in the highest release volume despite the second lowest release rates.

**Table 3:** Alberta historical NCM release rates by location and fiscal year.

FY	Calgary	Cold Lake (and Medley)	Edmonton	Suffield	Wainwright
	a.p. = 68	a.p. = 1,574	a.p. = 4,012	a.p. = 103	a.p. = 628
04/05	13.2%	6.4%	9.0%	8.2%	16.6%
05/06	10.1%	8.9%	9.2%	9.1%	10.8%
06/07	26.3%	9.9%	10.4%	15.5%	15.8%
07/08	18.3%	9.6%	13.4%	10.6%	16.4%
08/09	78.8%	7.6%	11.1%	13.5%	9.2%
09/10	7.1%	5.6%	7.4%	7.0%	11.3%
10/11	8.1%	5.5%	8.7%	9.7%	14.0%
11/12	5.5%	6.1%	7.6%	21.1%	7.9%
12/13	2.9%	8.5%	11.9%	7.6%	8.7%
13/14	18.4%	12.9%	12.4%	8.7%	18.6%

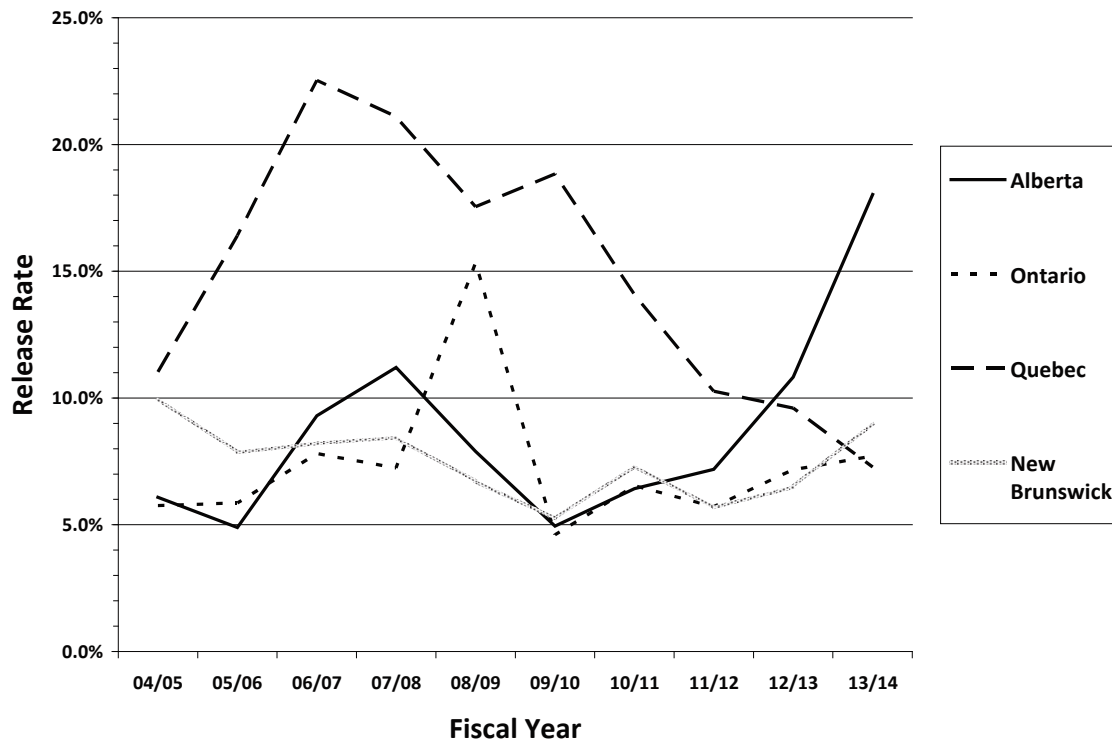
The next step in the analysis was to determine whether the increased release rates were common among all of the military occupations in Alberta. Although many different occupations are present in Alberta, the eight most populous occupations (shown in Table 4) serve as a representative sample: they account for approximately 61% of all NCMs and 67% of all NCM releases in the province. Release rates that have increased for two or more consecutive years (regardless of size) are shaded in Table 4. Observe that most of the occupations showed substantial increases in release rates in the last two to three years. Most notably, the combat engineer (CBT ENGR) occupation in Alberta showed a consistent year-over-year increase in release rates since FY 09/10. Conversely, preliminary FY 13/14 release rates for the resource management support (RMS) clerk and vehicle technician (VEH TECH) occupations are lower than those for FY 12/13. Therefore, as was found in the investigation of release rate differences between provinces and between locations in Alberta, occupation release rates within Alberta also tend to exhibit different patterns.

**Table 4:** Comparison of NCM historical release rates in Alberta by occupation.

FY	AVN TECH	CBT ENGR	CRMN	INFMN	MSE OP	RMS CLERK	SUP TECH	VEH TECH
	a.p. = 562	a.p. = 377	a.p. = 428	a.p. = 1,274	a.p. = 337	a.p. = 278	a.p. = 416	a.p. = 433
04/05	6.7%	6.1%	9.1%	11.1%	8.8%	5.8%	7.2%	6.6%
05/06	10.6%	4.9%	10.2%	12.6%	12.3%	8.4%	8.2%	7.3%
06/07	11.7%	9.3%	11.4%	14.1%	9.2%	7.4%	9.1%	11.9%
07/08	10.0%	11.2%	16.1%	14.6%	14.6%	9.5%	13.1%	14.0%
08/09	9.4%	7.9%	10.5%	9.4%	12.0%	8.9%	9.4%	18.8%
09/10	5.9%	5.0%	8.3%	8.5%	10.6%	7.4%	4.8%	7.8%
10/11	4.5%	6.4%	8.1%	13.9%	9.0%	9.2%	5.1%	9.3%
11/12	6.4%	7.2%	10.1%	9.0%	6.0%	9.6%	4.5%	6.3%
12/13	8.5%	10.8%	15.9%	15.8%	11.1%	8.6%	7.1%	10.8%
13/14	11.2%	18.0%	19.0%	19.6%	11.1%	6.5%	8.3%	10.3%

Although Table 4 shows that many occupations in Alberta have experienced increasing yearly release rates, it was also important to verify whether similar trends are found in other provinces. Figure 2 presents the combat engineer occupation release rates as an example. This occupation was chosen here because it has had a pronounced increase in release rates over the past few years. Note that release rates for the combat engineer occupation are only shown for four provinces, because the combat engineer population in the other provinces are too small to provide valuable statistics for comparison. As can be observed, Ontario's and New Brunswick's historical release rates for this occupation have also increased since FY 09/10 albeit not consistently and not to as large of an extent as Alberta. Interestingly, combat engineer release rates in Quebec have been decreasing

since FY 09/10. Therefore, it can be concluded that the rapid increase in combat engineer release rates since FY 09/10 is predominantly occurring in Alberta.



**Figure 2:** Comparison of combat engineer occupation historical release rates by province.

Next, YOS and age demographics were examined to determine if they can provide insight into the increasing Alberta release rates. Although both age and YOS demographics were analysed, this paper only presents the YOS figures because there were no major trends observed with the age. Table 5 presents the historical YOS at release for members by province. Unlike in the previous tables, two or more years of consecutive *decreases* are shaded in Tables 5 and 6 in order to highlight trends. Overall, the average YOS at release values were fairly stable within a province until FY 09/10 for most provinces. Despite being relatively stable until FY 09/10, however, YOS at release in Alberta over the last three years is nevertheless 2 to 3 years lower on average than it was in the FY 04/05 to 09/10 period. Further, YOS at release in FY 13/14 for Alberta is lower than it is in all of the other provinces except Quebec. This suggests that CAF members in Alberta tend to release from the military earlier in their careers and/or that there are more members at lower YOS who are available to release. Finally, YOS at release appears to have increased in the last year for all of the provinces examined in this study. In the case of Alberta, however, the increase was much smaller than that of the other provinces.

**Table 5:** Comparison of average NCM YOS at release by province.

FY	British Columbia	Alberta	Manitoba	Ontario	Quebec	New Brunswick	Nova Scotia
	a.p. = 4,399	a.p. = 6,398	a.p. = 2,184	a.p. = 15,972	a.p. = 8,896	a.p. = 3,654	a.p. = 6,717
04/05	16.5	14.6	17.1	18.2	10.1	15.0	19.2
05/06	17.6	14.5	16.9	18.8	9.3	13.2	19.2
06/07	17.3	15.5	18.8	17.9	8.2	14.9	20.1
07/08	15.8	13.5	15.0	17.4	5.2	11.3	19.5
08/09	14.0	13.0	13.7	14.6	8.9	11.7	17.1
09/10	17.0	13.3	17.4	16.0	5.1	13.7	20.5
10/11	15.4	10.6	15.1	16.1	6.7	8.9	18.6
11/12	15.9	12.0	13.8	15.5	9.5	11.1	19.6
12/13	14.3	11.1	13.1	16.6	9.8	11.3	17.5
13/14	15.7	11.2	17.9	18.6	11.0	13.1	18.1

Table 6 provides the average YOS at release for just the combat engineer occupation. The key finding in this table is that the combat engineer average YOS at release has been decreasing in Alberta since FY 09/10, but this continuous decline is not observed in the other three provinces listed. In fact, combat engineer occupation average YOS at release in FY 13/14 is approximately half of that in FY 09/10. Further, by comparing Tables 5 and 6, it can be deduced that not all occupations in Alberta have experienced increasingly lower YOS at release, as is the case with the combat engineers. Specifically, Alberta's average YOS at release for all the occupations has increased from FY 10/11 to FY 11/12, and from there it has remained fairly steady in FY12/13 and FY 13/14. This observation indicates that there are other occupations in Alberta that are countering the impact of the decreasing YOS at release caused by the combat engineer occupation.

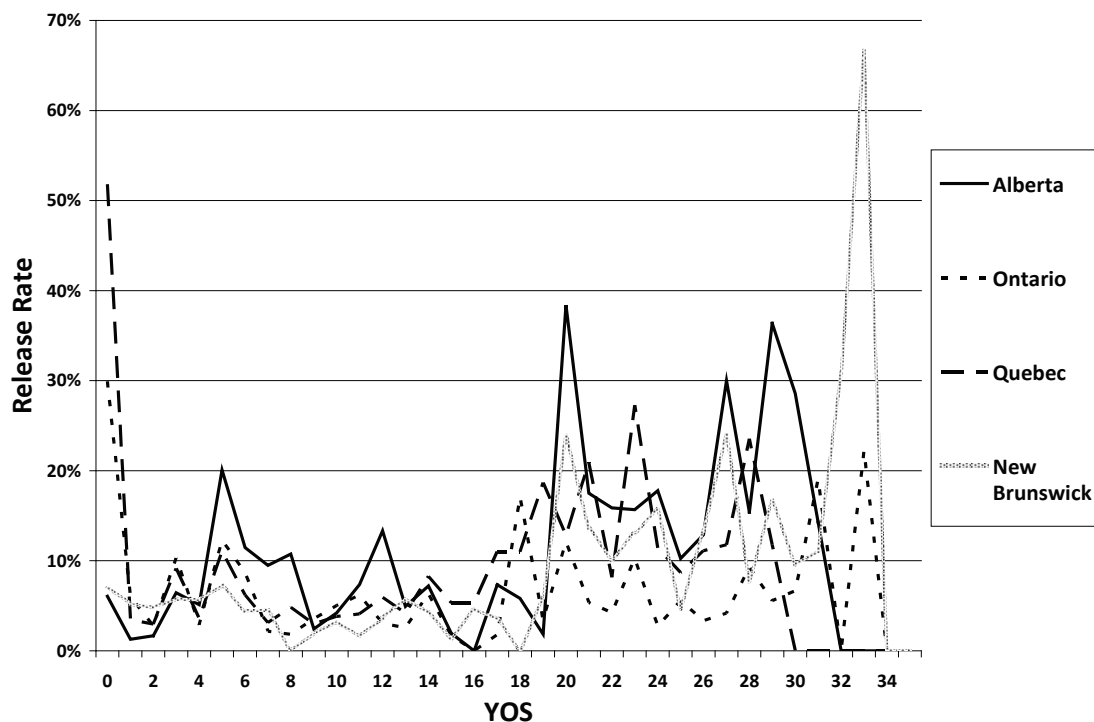
In addition to Alberta, there is significant variability in combat engineer YOS at release between years in each province. Combat engineer average YOS at release in Ontario and New Brunswick were also lower in FY 13/14 than in FY 09/10, but the difference was less pronounced than in the case of Alberta. Similarly, Quebec average YOS at release in FY 13/14 is less than half of that in FY 04/05. However, despite pronounced fluctuations, the general trend for all the provinces is that of decreasing average YOS release, now ranging from 4 to 8 YOS. However, only Alberta has been showing a continuous decrease in average YOS at release since FY 09/10 where as the other three provinces have experienced both increases and decreases.

**Table 6:** Comparison of combat engineer average YOS at release by province.

FY	Alberta	Ontario	Quebec	New Brunswick
	a.p. = 401	a.p. = 395	a.p. = 411	a.p. = 511
04/05	12.9	12.7	9.0	10.3
05/06	9.9	11.2	3.6	4.8
06/07	13.3	5.2	2.6	6.4
07/08	12.5	11.6	1.5	4.3
08/09	10.3	5.6	6.2	12.0
09/10	12.5	9.7	2.7	6.9
10/11	8.5	11.0	3.6	4.5
11/12	7.5	5.4	5.1	4.9
12/13	7.2	6.2	5.8	9.9
13/14	6.4	7.8	4.1	6.5

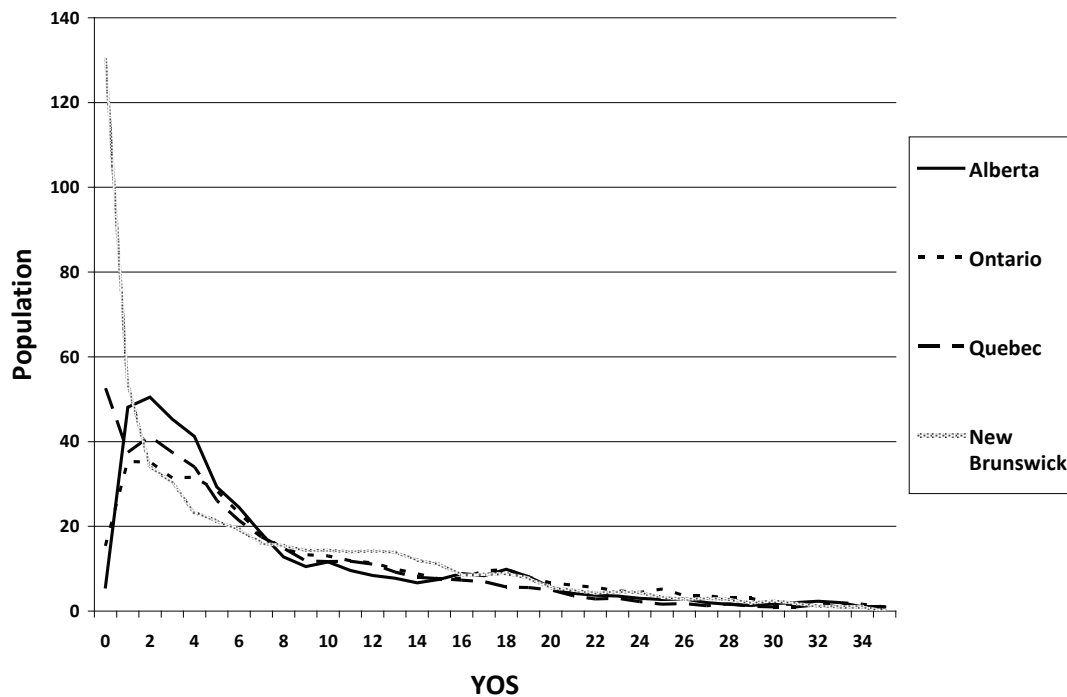
The historical release rates by YOS and the YOS profiles for combat engineer populations in Alberta, Ontario, Quebec, and New Brunswick were also compared. This approach provides a means to investigate whether the increased release rates are caused by more members reaching the YOS points with high release rates in recent years, by different release behaviours, or a combination of both. For example, if the combat engineer population in Alberta was found to have proportionally more members at or near the YOS points which

typically experience higher release rates, it would be expected that more members would have left the occupation in that province. Figure 3 shows an eight-year average YOS at release profile for the combat engineer occupation in the four provinces where the occupation is mainly located. The information presented in this figure leads to a better understanding of when in their careers members might chose to leave this occupation in a particular province. As was stated earlier in this paper, Quebec has the highest release rate at 0 YOS due to the release of untrained members during BMQ. Although Ontario release rates at 0 YOS are also high, Ontario combat engineer population at 0 YOS is small thus resulting in large release rate fluctuations. Alberta actually has the lowest release rates at the 0 to 4 YOS points. However, Alberta does show the highest release rates at several intervals, including 5 to 8 YOS, 11 to 12 YOS, 20 YOS, and 27 YOS. The fact that Alberta release rates are the highest at several YOS points is not unexpected because, as Figure 2 showed, Alberta's overall release rates for combat engineers have been increasing rapidly since FY 09/10 and have surpassed those of the other three provinces.



**Figure 3:** Comparison of combat engineer historical attrition rates by YOS and province.

Figure 4 shows the ten-year average YOS profiles of the combat engineer populations in the four provinces. Observe that the relative proportions of members at each of the YOS points differ (predominantly) between the provinces at 0 to 5 YOS and to a lesser extent at 10 to 19 YOS. Specifically, other than at 0 and 1 YOS, Alberta, on average, has a greater population of members with lower YOS than the other three provinces. Further, Figure 3 already showed that Alberta's historical release rates are also higher than those of the other three provinces from 5 to 8 YOS. Therefore, if a large number of Alberta combat engineers at 1 to 5 YOS points shown in Figure 4 release once they reach the points of high attrition shown in Figure 3, then this would cause more combat engineers to leave their careers than in the other three provinces. While Ontario and Quebec have much higher release rates than Alberta at 0 YOS, there are also not many members at this YOS to have much impact on the average provincial YOS at release rates for combat engineers.



**Figure 4:** Comparison of combat engineer average YOS profiles by province.

#### 4. Discussion

This study's findings augment DGMPPRA's 2013 study by updating attrition rates by province with more recent data and in more detail. Results suggest that release rates in Alberta have increased in the last two years. In fact, all the provinces have been experiencing increased release rates in the last two to three years. However, all provinces, including Alberta, have also experienced years of increasing and decreasing release rates. Therefore, it is difficult to determine whether the recent increases in release rates will continue or whether the trend will eventually reverse. For example, release rates in Alberta increased from FY 04/05 to 07/08, but then declined from 07/08 to 09/10.

This study further shows that not all locations across Alberta have experienced an equal increase in release rates. While Cold Lake (and Medley) has experienced a pronounced increase in the yearly release rates for the past three years, this was not found to be the typical case for the other locations. Edmonton and Wainwright release rates also increased, but only in the last two years. On the other hand, Calgary and Suffield rates decreased from FY 11/12 to 12/13 and then increased only in the last year. There is therefore value in examining each location individually, although one should factor in the size of the population and the YOS profile when comparing the changes or differences in release rates.

With respect to the nature of release rates within the various military occupations in Alberta, it was found that most of the occupations have experienced increased release rates in the last three to five years. However, there are instances of occupations, such as the combat engineers, where the release rates have increased far more rapidly than in other occupations. Further, RMS clerk occupation release rates actually decreased over the last two years. Overall, historical release rates by occupation fluctuated significantly, requiring caution when drawing conclusions based on only a few years of data. Therefore, while the overall Alberta release rates have increased, there is evidence to suggest that release rates in Alberta are not increasing across all occupations.

Finally, analysis of the combat engineer population average YOS at release combined with the release rates by YOS indicate that more members in Alberta were releasing at an early stage in their careers (between 5 to 12 YOS) than in the other provinces. Demographic analyses showed that Alberta has historically had a large combat engineers population at 1 to 5 YOS and high release rates at 5 to 8 YOS. Future studies will have to verify whether the same findings apply to the other occupations in Alberta.



## **5. Limitations**

A number of limitations should be considered when evaluating these findings. First, this study was created before the end of FY 13/14; it was therefore assumed that release volumes for the remainder of the year could be forecast from historical release patterns during the final two months of previous FYs. However, it is possible that disproportionately fewer or more members released in the last two months of the FY than expected, which would affect the final release volumes for FY 13/14. However, data will be updated and quality will improve over time as the release and population figures are updated in the HRMS database. Second, the methodology used in this study is not able to explain the underlying motivations for voluntary releases from the CAF. This area of research could be addressed through separate studies, such as an analysis of responses from the CAF Exit Survey, which polls voluntarily releasing members regarding their reasons for release. Third, some of the occupations present in Alberta were not analysed because there were not enough members in their populations and/or historical releases to reliably uncover trends. Fourth, this study did not differentiate between trained and untrained (recruits) members. As a result, it is possible that different findings could emerge if recruits were not included in the analysis or if the study focused only on recruits. Fifth, most of the tables and figures presented in this document illustrate yearly fluctuations in release volumes and/or rates. Therefore, the release rates should be reassessed periodically to identify whether rates have changed. Finally, an analysis of the statistical significance of increases/decreases observed in the reported annual attrition rates was not included in this study.

## **6. Conclusion**

Overall, this study found that NCM release rates in Alberta and other examined provinces have increased over the last two years, but the effect is more pronounced in Alberta. However, the release rates have not increased universally across Alberta over this timeframe. First, while all of the locations in Alberta showed increased release rates since FY 12/13, the magnitude of the increase varied across locations. Next, although occupations such as combat engineers showed pronounced increases in yearly release rates in Alberta, this was not a typical observation because some occupations experienced much smaller increases or even decreased release rates. In fact, the RMS clerk occupation in Alberta experienced large decreases in the yearly release rates. Finally, demographic analyses of the combat engineer population suggest that members in Alberta were releasing earlier in their careers than in the other provinces. Future studies will have to verify whether the same findings apply to the other occupations in Alberta.

Therefore, although Alberta has experienced higher release rates than the other provinces, release rates appear to vary by locations and military occupations on a case-by-case basis. Additional analysis into the underlying reasons for these differences should allow for better management of those CAF locations and/or military occupations experiencing high release rates.

## **References**

- Fang, M. and Bender, P. (2009) *Canadian Forces Attrition Analysis by Region*. INFORMS-MAS Conference.
- Latchman, S. (2013) *Recent and ongoing analyses of attrition by location*. Briefing Note for DGMPPRA. Ottawa, Ontario: Director General Military Personnel Research & Analysis.
- Okazawa, S. (2007) *Measuring Attrition Rates and Forecasting Attrition Volume*. CORA TM 2007-02. Defence R&D Canada – CORA.
- Otis N. and Straver, M. (2008) *Review of Attrition and Retention Research for the Canadian Forces*. CORA TM 2008-030. Defence R&D Canada – CORA.

# Importance, Use and Organizational Assignment of Web Analytics for Social Economy Organizations in Austria – An initial empirical study

Claudia Brauer<sup>1,2</sup>, Stefan Bauer<sup>2</sup> and Julia Hauptmann<sup>3</sup>

<sup>1</sup>Management Center Innsbruck, Innsbruck, Austria

<sup>2</sup>Institute for Information Management and Control, Vienna University of Economics and Business, Vienna, Austria

<sup>3</sup>FH Campus Wien, Vienna, Austria

[claudia.brauer@mci.edu](mailto:claudia.brauer@mci.edu)

[stefan.bauer@wu.ac.at](mailto:stefan.bauer@wu.ac.at)

**Abstract** In recent years, Web Analytics has increasingly gained importance due to the necessity to measure the success of Websites and online social networks. Especially for social economy organizations, Web Analytics offers a great variety of useful applications to support decision makers. There are less research results regarding the importance, the use and the organizational assignment of Web Analytics in social economy organizations in the German-speaking, as well as other European countries. Therefore, this paper investigates the importance, use and organizational assignment of Web Analytics in social economy organisations in Austria. We conducted an online survey among 272 social economy organizations in Austria. The presented initial research study revealed that Web Analytics in social economy organizations in Austria is due to different reasons still in its infancy.

**Keywords:** Big data, web analytics, social economy organizations, Austria, quantitative research method, online survey, internet data analysis

---

## 1. Introduction

In recent years, social economy organizations (according to Defourny 2004) have recognized the importance of online marketing and social media marketing to achieve their organizational goals. Due to the increasing importance of an organization's Website and online social networks for communication and marketing purposes for social economy organizations and due to decreasing public funds, the performance measurement of the organization's Website and online social networks for social economy organizations has gained more importance (Lovejoy/Saxton 2012; Waters et al. 2009). The success of the organization's Website and online social networks can be analysed with Web Analytics software, e.g. Google Analytics. With a Web Analytics software it is possible to analyse in real-time the information search and purchasing behaviour of Website visitors based on different Web metrics e.g. page impressions, to make informed strategic decisions (Calero et al. 2005; Palmer 2002). Therefore Web Analytics is, according to Chen et al. 2012, an expression of big data and support different decisions makers within the organization. Some authors have already concluded that Web Analytics has been developed to a strategic management tool, which offers a lot of applications (Agarwal and Venkatesh 2002; Delone and McLean 2004; Molla and Licker 2002; Pakala et al. 2012; Sterne 2002). Currently, there exists a great variety of Web Analytics software, which are intensively used by companies (Park et al 2010; Nakatani and Chuang 2010). A research study from Forrester Research (2011) confirmed that for 70% of the surveyed companies Web Analytics plays an important role in the organization. Zumstein et al. (2012) revealed in a quantitative research study that three-quarters of all SMEs in the German-speaking countries already use Web Analytics for corporate management, but only 1% of NPOs/NGOs is using Web Analytics. There are less research results regarding the use and the importance of Web Analytics in social economy organizations in the German-speaking, as well as in other European countries. Against this background, this paper examines the importance and use of Web Analytics for social economy organizations as an example from Austria in order to describe the current situation in social economy organisations. Specifically, we address the following research questions:

- How diffused and aware is Web Analytics in social economy organizations in Austria?
- What are the objectives of Web Analytics in social economy organizations in Austria?
- In which departments is Web Analytics assigned to social economy organizations in Austria?
- Which Web Analytics software is used by social economy organizations in Austria?

- Which Web metrics are used by social economy organizations in Austria to analyse the performance of the Website and other online social networks?

Therefore, the research aim of this article is to analyse the current integration (e.g. importance, applications, organizational assignment, applied Web Analytics software, problems concerning Web Analytics) of Web Analytics in social economy organizations in Austria. The research study represents an initial investigation concerning the current use of Web Analytics in social economy organizations in German-speaking countries and in the European region generally. In the next section, theoretical considerations about the definition, objectives and applications, organizational assignment of Web Analytics based on scientific literature are presented. Web Analytics, as the following research study demonstrates, has not been scientifically investigated from the perspective of social economy organizations; therefore only literature from the field of Web Analytics and internet data analysis is considered. The third section describes our research design. Based on the description of research method, the research results are presented in detail. Finally, the re-search findings are discussed and new research areas for social economy organisations are explained.

## **2. Theory about Web Analytics**

### **2.1 Definition of Web Analytics**

Despite the increasing importance of Web Analytics and long-standing existence age (since 1995) of Web Analytics, there is still no common understanding of Web Analytics exists in science and practice (Calero et al. 2005; Chen et al. 2012; Palmer 2002). A great variety of definitions from different disciplines, synonymous terms, e.g. internet data analysis, Web measurement, Website analysis, Website monitoring, Website measurement, Web (site) statistics and Web tracking, as well as different spellings can be evidenced in the scientific literature. Thus, the term of Web Analytics has taken a multifaceted and interdisciplinary development, which currently offers room for interpretation within the different disciplines. Furthermore, in both science and practice, as can be seen in Table 1, a different understanding of Web Analytics still exists. As can be seen in Table 1, Erben (2002) sees Web Analytics from the management accounting and controlling perspective. Weischedel et al. (2005), Waisberg and Kaushik (2009) and Digital Analytics Association (2006) consider Web Analytics from the marketing point of view. Kumar et al. (2012) and Gartner Group (2001) are more focused on informatics perspective of Web Analytics.

### **2.2 Objectives and applications of Web Analytics**

The objectives and applications of Web Analytics are closely linked to the objectives of e-Business. The original objective of Web Analytics was to improve the usability of the Website (Weischedel and Huizingh 2006). The objectives of e-business are manifold, e.g. cost reduction and increasing customer loyalty, and depend on organizational objectives. Therefore, Web Analytics can be used to measure and analyse a variety of different organizational objectives and to determine the achievement of these goals. Accordingly, Web Analytics offers, due to the technical developments in recent years, e.g. as mouse tracking, versatile applications which can be usefully applied for strategic corporate management. Scientific articles which are mainly investigating the objectives and the applications of Web Analytics appear to be scarce (Park et al. 2010). Currently, only few scientific articles explicitly investigate the objectives and use of Web Analytics software in companies (Park et al. 2010; Pakala et al. 2012; Devaraj et al. 2002). Research studies about the objectives and applications of Web Analytics in social economy organizations, as already mentioned, are not apparent in the scientific literature. Often in scientific publications, technical requirements for Web Analytics software are investigated and the goals and the applications of Web Analytics are not clearly delineated (Alpar et al. 2001; Chen et al. 2012; Wade and Nevo 2005; Pakala et al. 2012). In recent years, several new Web Analytics applications due to development of information and communication technologies, especially in the areas of user identification, data collection, and the online marketing, have emerged. Despite the variety of objectives of Web Analytics, as can be seen in Figure 1, Web Analytics can be categorized as follows:

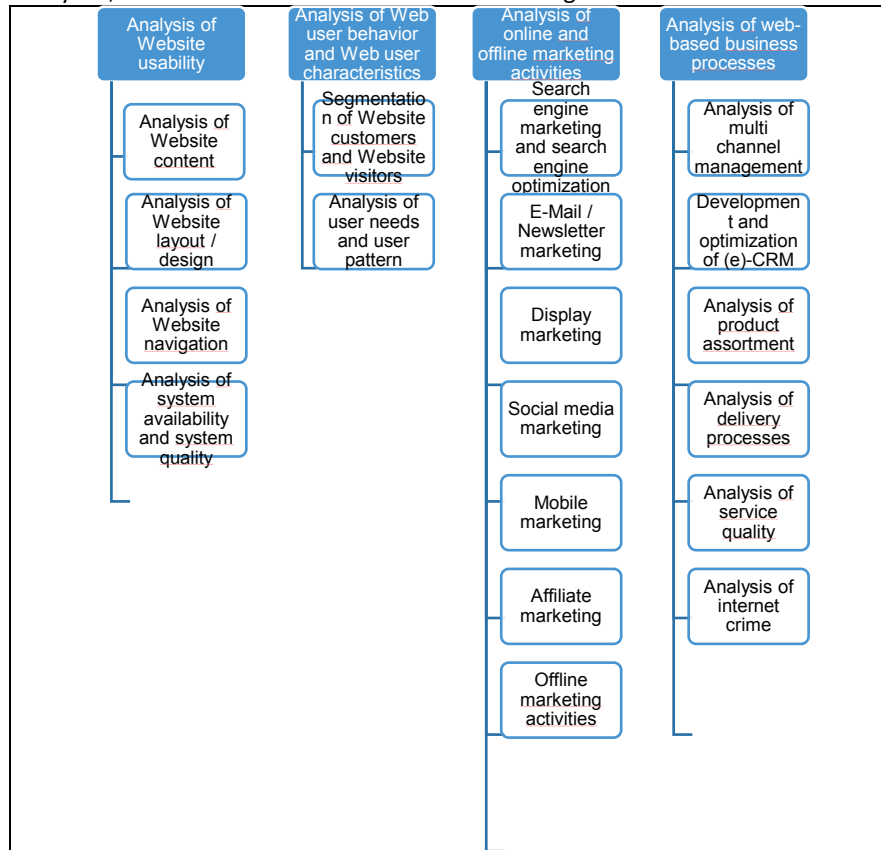
Author	Definition	Target area	Discipline
Gartner Group 2001	Web Analytics uses a variety of data and sources to evaluate Web site performance and visitor experience, potentially including usage levels and patterns at both an individual and aggregate level. Data and sources may include click-stream data from Web server log, Web transaction data, submitted data from input fields on the Web site and data in the repository. The goals are to improve site performance, both from a technical and a content perspective, enhance visitor experience (and thus loyalty), contribute to overall understanding of customers and channels and identify opportunities and risks.	Practice	Business informatics
Erben 2002, p. 59	In general, this term includes all those activities in management and accounting that are connected with planning, controlling and monitoring the Internet activities of a company.	Science	Management Accounting / Controlling
Weischedel et al. 2005, p. 2	Web Analytics as the monitoring and reporting of Web site usage so that enterprises can better understand the complex interactions between Web site visitor actions and Web site offers, as well as leverage insight to optimize the site for increased customer loyalty and sales.	Science	E-Business/ E-Marketing
Web Analytics Association 2006 (= Digital Analytics Association)	Web Analytics is the measurement, collection, analysis and reporting of internet data for the purposes of understanding and optimizing Web usage. Web Analytics is the objective tracking, collection, measurement, reporting and analysis of quantitative Internet data to optimize Websites and Web marketing initiatives.	Association	Web Analytics E-Business / E-Marketing
Waisberg and Kaushik 2009, n.p.	„Web Analytics deals with the methods for measurement, data collection, data analysis and providing the related feedback on internet for the motive of understanding behaviour of the customer using Website. The benefit of studying behaviour of the customer leads to optimize the usage of Web site.“	Practice	E-Business / E-Marketing
Kumar et al. 2012, p. 966	“As per nomenclature in Web Analytics is the art and science of improving the customer’s Website experience. This is a science, because it uses statistics, data mining techniques, report summaries, operations and a methodological process to process. It can be said as art because to improve Websites there requires a deep level of creativity, imagination, analysis, balancing user-centric design, promotions, content, images and more.“	Science	Informatics

**Table 1.** Definitions of Web Analytics

### 3. Research Design

Based on research study of Zumstein et al. (2012) about Web Analytics for profit organizations in all branches, we developed an online survey that was sent to all social economy organizations in Austria. Consequently, a Website was a prerequisite for participation on online survey. We conducted a pre-test in advance; therefore, we sent the online survey to 20 persons in different social economy organizations in Austria. The online survey includes four questions about demographic characteristics of survey participants (e.g. state, organizational form, number of employees, field of practice) and 18 questions about use and knowledge of Web Analytics (e.g. organizational assignment, objectives, period of use, reason of use etc.). The selection of respondents is based on list that is created and maintained by Austrian Federal Ministry of Social Affairs and Consumer Protection. The research study only includes social economy organizations that are exclusively working in the non-profit sector; this includes religious institutions (e.g. church-run nursing homes), private organizations or semi-public social economy organizations. Based on the list, the online questionnaire was sent to 1.293 social organizations in Austria by email. Due to wrong email addresses and overloaded mailboxes 1.176 social organizations have received the online survey. The online survey was conducted in the period from 14/05/2013 to 14/06/2013. Most online questionnaires were completed or clicked in the first 10 to 14 days. A total of 374 online questionnaires were returned by social economy organizations, but only 272 surveys were fully completed. This corresponds to a response rate of 31.8%. The regional distribution of the participating social economy organizations is similar to the average distribution of social organizations in Austria. 31% of social economy organizations are located in Vienna, 19.4% in Lower Austria and 16.2% in Upper Austria. The participation rates from all other states of Austria were between 14.7% and 8.5%. Additionally, we asked the

social economy organizations to categorize themselves based on the number of employees. 57.1% of the social economy organizations are small organisations (between 1-49 employees). 28.5% are medium-sized (between 50-250 employees) and 14.4% are large organizations (more than 251 employees). For data analysis we used SPSS version 20, and we applied different univariate statistical methods and two/multivariate statistical methods. Unfortunately, the results of the research demonstrate that only few social economy organizations are already applying Web Analytics in their organization. Due to the small number of social organizations that are using Web Analytics, multivariate methods lack of statistical significance.



**Figure 1.** Applications of Web Analytics (Alpar et al. 2001; Wade and Nevo 2005; Pakala et al. 2012; Chen et al. 2012; Park et al. 2010; Weischedel et al. 2005; Nakatani and Chuang 2010)

## 4. Research Results and Discussion

### 4.1 Awareness and Application of Web Analytics in Social Economy Organizations

The research study shows that the majority of Austrian social economy organizations are aware of the term “Web Analytics”, but only a few of them are practicing it. 53% of the investigated social economy organizations have a specific knowledge about Web Analytics, but only 22% (74) of them apply Web Analytics to monitor and analyse their Web services. Furthermore, 14% of the social economy organizations, that presently do not use Web Analytics, want to implement Web Analytics software in the future. In contrast to knowledge about the term Web Analytics itself, only 31% of the respondents have knowledge about applications of Web Analytics. Compared to profit organizations, the social economy organizations in Austria are using Web Analytics less than other industries branches (Zumstein et al. 2012).

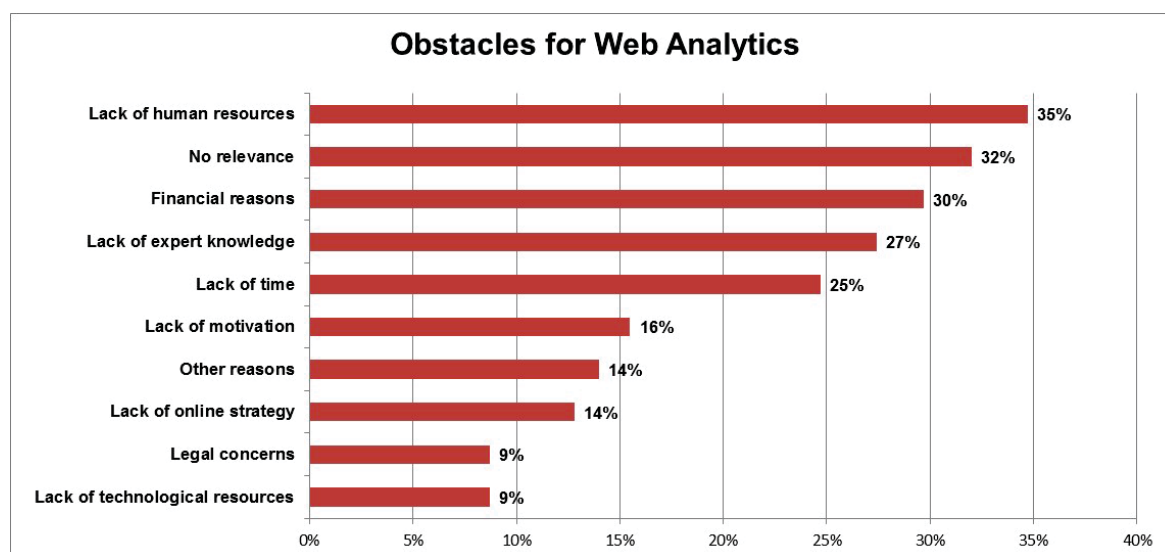
Regarding the size of organization, we found that 45% of large social economy organizations have implemented Web Analytics within the organization. 24% medium-sized and 30% small social organizations are using Web Analytics to measure the performance of their Website. Our findings strongly indicate that the organizational size of the social economy organization is an indicator for the application of Web Analytics ( $\chi^2=21,800$ ;  $p=0,000$ ). Thus we assume that the more employees an organization has, the more human and financial resources can be provided for Web Analytics.

The social economy sector is divided into a great variety of different fields of practice. As can be seen in Table 2, Web Analytics is more common in the practice fields care of older people (20%), work, education and occupation (19%), children, young adult's care (19%) and health (19%) than in the fields of practice disability (5%), addiction (4%), and family, partner, and single parents (9%). Inexplicably, no social economy organization in the field of practice migration is using Web Analytics. The correlation analysis showed that the field of practice has no influence on the application of Web Analytics in social economy organizations.

Field of Practice	Application of Web Analytics	
	Numbers	Percentage
Care of older people	15	20%
Work, education and occupation	14	19%
Children / Young adults care	14	19%
Health	14	19%
Family / Partner / Single parents	7	9%
Disability	4	5%
Addiction	3	4%
Housing and residential environment	2	3%
Delinquency	1	1%
Migration	0	0%
<b>Overall</b>	<b>74</b>	<b>100%</b>

**Table 2.** Fields of Practice (N=374)

There are several obstacles for social economy organizations in implementing Web Analytics in their organizations. As can be seen from Figure 2, the most notable obstacles are the lack of human (35%) and financial (30%) resources. Interestingly, one third of the respondents answered that currently Web Analytics has less or no relevance for them. Further, 27% replied that they have a lack of technical and expert knowledge in the field of Web Analytics. 25% of the social economy organizations have indicated a lack of time and 16% a lack of motivation. Some respondents (14%) explained that their social economy organizations have no strategy for their online presence. Nearly 9% of the respondents mentioned that they either have legal and privacy concerns about Web Analytics or have a lack of technological resources.



**Figure 2.** Obstacles for Web Analytics (N=31)

Previous research has underpinned the importance of external consultants for Web Analytics (Zumstein et al. 2012). 11% of the social economy organizations have engaged external consultants before they implemented a Web Analytics software, and 22% have consulted external experts. The external consultancy supported the



social economy organizations by developing strategic objectives for their Website and implementing suitable Web Analytics software.

#### **4.2 Objectives and Applications of Web Analytics in Social Economy Organizations**

Social organizations pursue a different of objectives with their Web presence. The objectives of the Website for social economy organizations in Austria are to provide information (83.8%), to improve communication (71%) with the different stakeholders and to support brand awareness (77.5%). Further, social economy organizations use their Website to improve customer satisfaction and customer loyalty (64.3%) and to acquire new customers (61.3%). Some social economy organizations (51.6%) use their Website for personnel issues, e.g. online applications. The applications of Web Analytics are closely linked with objectives of the Website. All social economy organizations reported that they assess the usability and optimization potential of their Website. 80.4% are using Web Analytics to analyse the content quality of their Website, and nearly 70% are monitoring the Website user behaviour and Website user characteristics. 65.2% of the social organizations are regularly investigating the navigational elements, and 52.2% of the social economy organizations analyse the layout and design of their Website using Web Analytics software. 63% are evaluating the service quality. Interestingly, 54.4% of the social economy organizations are applying Web Analytics to evaluate their multichannel marketing activities and 45.7% to evaluate their search marketing activities and search engine optimization efforts. More social economy organizations (41.3%) analyse rather their social media marketing activities than their online marketing activities (34.8%) with the support of an adequate software. 34.7% of the social economy organizations are applying Web Analytics to development their e-CRM.

#### **4.3 Organizational Assignment of Web Analytics in Social Economy Organizations**

Web Analytics is an interdisciplinary working task, which needs high communication and analytic skills (Chen et al. 2012; Kumar et al. 2012). As mentioned before, Zumstein et al. (2012) have revealed that most profit organizations have integrated Web Analytics into their marketing department. The majority of social organizations in Austria also indicate that Web Analytics is a working task of their marketing department (33%). 24% of the social economy organizations assign Web Analytics to the corporate management. The assignment of Web Analytics to corporate management emphasizes the importance of Web Analytics as a strategic management tool. Further, 17% of the social economy organizations in Austria have Web Analytics assigned to their IT departments and 26% have integrated Web Analytics to other departments, e.g. controlling.

#### **4.4 Web Metrics in Social Economy Organizations**

Web Analytics software supports social economy organizations to create internal reports about the performance of their Web presence for different stakeholders. These reports include a multiplicity of Web metrics. The selected Web metrics are affected by the objectives of the Website (Calero et al. 2005). As can be seen in Figure 3, nearly all social economy organizations are using standard Web metrics, e.g. page impressions. Content metrics are analysed and monitored by 90% of the social economy organizations. 84% monitor the visitor behaviour and traffic sources, therefore both Web metrics seem to be important for social economy organizations. Three quarters of the participants investigate Website visitor characteristics and 42% determine Web 2.0 metrics.

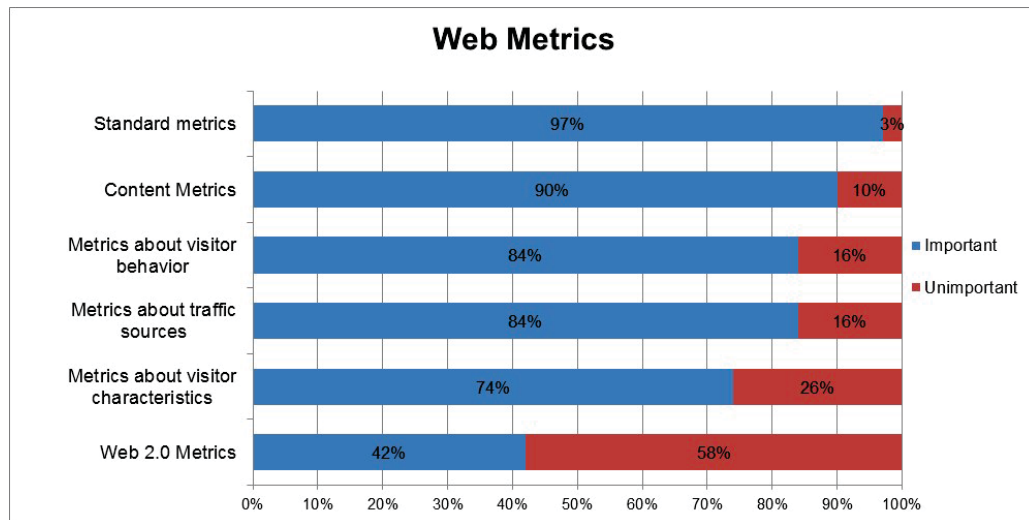


Figure 3. Web Metrics of Social Economy Organizations (N=47)

#### 4.5 Web Analytics Software in Social Economy Organizations

As previously mentioned, the release of Google Analytics in 2006 (before Urchin) and Piwik in 2008 marked the beginning of Web Analytics in organizations. 20% of the respondents have been using Web Analytics system for more than 5 years. 17% of the researched social economy organizations have implemented Web Analytics software since 4 to 5 years and 48% of them have been applying Web Analytics software since the last 2 to 3 years. Only 15% have been using Web Analytics software less than one year. As can be seen in Figure 4, nearly 60% of the social economy organizations use Google Analytics to analyse their Website. Only a few social economy organizations use AWStats (9.7%), etracker (3.2%), Open Web Analytics (3.2%) and AT Internet (3.2%). The majority of the above-named Web Analytics software offers less analysis possibilities than other Web Analytics software, e.g. Omniture, SAS. The Web Analytics software that is used can be considered as indicator for the development of Web Analytics within the organization (Pakala et al. 2012). Therefore, social economy organizations can be categorized as beginners in the field of Web Analytics. Despite their low level of development regarding experience and knowledge in the field of Web Analytics more than 90% of the social economy organizations are currently satisfied with their Web Analytics software.

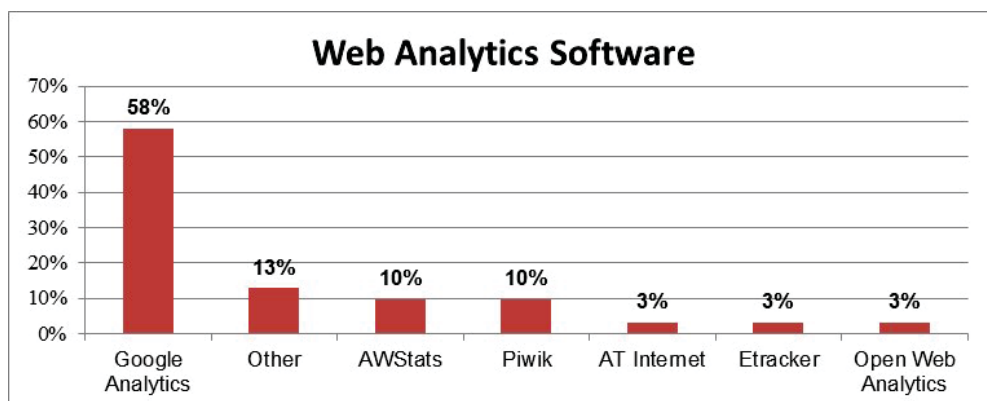


Figure 4. Web Analytics Software (N=31)

#### 5. Conclusion

The present study revealed that 53% of social economy organizations in Austria recognise the term Web Analytics. Despite that, for one third of social economy organizations, Web Analytics has currently no relevance. Interestingly, a significant correlation between the level of awareness of Web Analytics and the knowledge of the specific applications was determined. Therefore, we concluded that social economy organizations are aware of Web Analytics are familiar with its applications in the organization.



Additionally, the research study demonstrates that only few social economy organizations in Austria (22%) are using Web Analytics software to measure the success of their Website, their online marketing activities and their online social networks. Hence, Web Analytics in social economy organizations is still in its infancy. This development is primarily caused by a lack of personnel, time and financial resources. Therefore, the use of Web Analytics in social economy organizations depends on the size of the organization. The data illustrates that the larger the organization based on the number of employees, the more often Web Analytics is applied in social economy organizations in Austria.

Social economy organizations in Austria consider the analysis of the Website (100%) as the greatest benefits of Web Analytics. Furthermore, the majority of organizations consider the analysis of Web-site content as major benefit of Web Analytics. Moreover, some social organizations have recognized the importance of Web Analytics to improve their search engine marketing and search engine ranking. Resulting from the objectives of the Website, standard Web metrics, content metrics and metrics about website user behaviour and traffic sources are the most important ones for social organizations.

As previously discussed, the social sector is characterized by a variety of organizational forms and by a variety of fields of practice. Almost all these are using Web Analytics software. For-profit social organizations are especially interested in Web Analytics and its different applications. Nearly half of the social economy organizations have been using Web Analytics software for two to three years to measure the success of their Website and their online marketing activities. According to Zumstein et al. (2012) social economy organizations, in comparison to for-profit organizations have introduced (5-6 years), Web Analytics later. 58% of social economy organizations in Austria have utilised Google Analytics to analyse their Website. The majority of social economy organizations in Austria has assigned Web Analytics to marketing department. The assignment of Web Analytics to the marketing department illustrates that Web Analytics is mainly applied to measure marketing objectives of social economy organizations and this assignment is similar to those in for profit organizations. The use of Web Analytics to analyse the performance of the Website and other online social networks leads to a competitive advantage for social organizations and will support them to develop a sustainable digital strategy in the future (Chen et al. 2012; Pakala et al. 2012)

The research study has revealed that the obstacles for social economy organizations in implementing Web Analytics are based on a lack of human resources, expert knowledge within the social organizations, clear Website objectives and concerns regarding data protection and privacy. In order to receive more detailed research results about the objectives of Web Analytics and use of different Web metrics in social organizations in Austria, in-depth qualitative studies are needed. Especially with respect to experience of Web Analytics software in NPOs expert interviews are required. A comparison concerning the use of Web Analytics in further countries would be interesting. Finally, more detailed research results about the success factors of implementing and using Web Analytics software in social economy organizations are necessary. Research studies about the importance of Web 2.0 metrics are interesting in social economy organizations from the science perspective.

## References

- Agarwal, R. and Venkatesh, V. (2002). Assessing a Firms's Web Presence: A heuristic evaluation procedure for the measurement of usability. *Information Systems Research*, 13 (2), 168-225.
- Alpar, P., Porembski, M., Pickerodt, S. (2001). Measuring the Efficiency of Web Site Traffic Generation. *International Journal of Electronic Commerce*, 6 (1), 53-74.
- Calero, C., Ruiz, J. and Piattini, M. (2005). Classifying web metrics using the web quality model. *Online Information Review*, 29 (3), 227-248.
- Chang, C.-J., Torkzadeh, G. and Dhillon, G. (2003). Re-examining the measurement models of success for Internet commerce. *Information & Management*, 41 (5), 577-584.
- Chen, H., Chiang, R. H. L. and Storey, V. (2012). Business intelligence and analytics from big data to big impact. *MIS Quarterly*, 36 (4), 1165-1188.
- Digital Analytics Association (2006). The Official DAA Definition of Web Analytics. Retrieved from [www.digitalanalyticsassociation.org](http://www.digitalanalyticsassociation.org)
- Defourny, J. (2004) From third sector to social enterprise, in *The emergence of social enterprise*, Bozaga, C., Defourny, J. Routledge Taylor and Francis Group, London and New York
- Delone, W. and McLean, E. (2004). Measuring e-commerce success: applying the Delone & McLean information systems success model. *International Journal of Electronic Commerce*, 9 (1), 31-47.

- Devaraj, S., Fan, M. and Kohli, R. (2002). Antecedents of B2C channel satisfaction and preference: Validating e-Commerce Metrics. *Information Systems Research*, 13 (3), 316-333.
- Erben, R. (2002). E-controlling: Ten challenges for controlling and management accounting. *The Journal of Corporate Accounting & Finance*, 13 (5), 55-62.
- Forrester Research (2011). The Forrester Wave: Web Analytics, Q4, 2011, Forrester Research Inc. Reproduction Prohibited, 3-6.
- Gartner Group (2001). E-business dimension model: clarity behind the hype.  
<http://gartnerweb.com/public/static/hotc/00093980.html>. Retrieved on 05.12.2013.
- Kumar, L., Singh, H. and Kaur, R. (2012). Web Analytics and Metrics: A Survey. In *Proceedings ICACCI 2012*, 966-971.
- Lovejoy, K. and Saxton, G. (2012). Information, Community and Action: How Non-Profit Organizations use Social Media. *Journal of Computer-Mediated Communication*, 17, 337-353.
- Molla, A., Licker, P. (2001). E-Commerce system success: An attempt to extend and respecify the Delone and MacLean model of IS success. *International Journal of Electronic Commerce*, 2(4), 131-141.
- Nakatani, K. and Chuang, T. T. (2010). A web analytics tool selection method: an analytical hierarchy process approach. *Internet Research*, 21 (2), 171-186.
- Pakkala, H., Presser, K., and Christensen, T. (2012). Using Google Analytics to measure visitor statistics: The case of food composition websites. *International Journal of Information Management*,  
<http://dx.doi.org/10.1016/j.ijnfomgt.2012.04.008> Retrieved on 05.12.2013.
- Palmer, J. (2002). Web site usability, design, and performance metrics. *Information Systems Research*, 13(2), 151-167
- Park, J., Kim, J.J. and Koh, J. (2010). Determinants of continuous usage intention in web analytics services. *Electronic Commerce Research and Application*, 9 (1), 61-72.
- Sterne, J. (2002). *Web metrics: Proven methods for measuring web site success*. Wiley & Sons, New York.
- Wade, M. and Nevo, S. (2005). Development and Validation of a Perceptual Instrument to Measure E-Commerce Performance. *International Journal of Electronic Commerce*, 10 (2), 123-146.
- Waisberg, D., and Kaushik, A. (2009). Web Analytics 2.0: Empowering Customer Centricity, *Search Engine Marketing Research Journal* 2 (1), [http://www.semj.org/documents/webanalytics2.0\\_SEMJvol2.pdf](http://www.semj.org/documents/webanalytics2.0_SEMJvol2.pdf). Retrieved on 05.12.2013.
- Waters, R. D., Burnett, E., Lamm, A. and Lucas, J. (2009). Engaging stakeholders through social networking: How nonprofit organizations are using Facebook. *Public Relations Review*, (35), 102-106.
- Weischedel, B. and Huizingh, E. (2006). Website Optimization with Web Metrics: A Case Study. In: *Proceedings of the 8th international conference on Electronic commerce*, Fredericton, New Brunswick, 463-470.
- Weischedel, B., Matear, S. and Deans, K. (2005). The use of E-metrics in strategic marketing decisions: a preliminary investigation. *International Journal of Internet Marketing and Advertising*, (2/1-2), 109-125.
- Zumstein, D. (2012): Web Analytics. Explorative Analyse zu Einsatz, Nutzen und Problemen. Dissertation. Freiburg, <http://ethesis.unifr.ch/theses/ZumsteinD.pdf?file=ZumsteinD.pdf> Retrieved on 05.12.2013.

# Foundations of Big Data and Analytics in Higher Education

**Ben Daniel and Russell Butson**

**Higher Education Development Centre , University of Otago, Dunedin, New Zealand**

[ben.daniel@otago.ac.nz](mailto:ben.daniel@otago.ac.nz)

**Abstract:** This paper contributes to our theoretical understanding of the role Big Data plays in addressing contemporary challenges institutions of higher education face. The paper draws upon emergent literature in Big Data and discusses ways to better utilise the growing data available from various sources within the institutions of higher education, to help understand the complexity of influences on student-related outcomes, teaching and the ‘what-if questions’ for research experimentation. The paper further presents opportunities and challenges associated with the implementation of Big Data analytics in higher education.

**Keywords:** Big Data, Higher Education, analytics, improved learning, improved teaching, improved decision-making

---

## 1. Introduction

Institutions of higher education worldwide are operating in a rapidly changing and competitive environment, driven by political, cultural, economic, and technological factors. The complexity of these changes has far reaching impacts that affect every aspect of provision ranging from national systems of education to the curriculum and student learning. There are also growing regulatory demands for transparency and accountability while the sector experiences declining support from government, business and the private sector. These radical transformations necessitate new ways of thinking, research theories and approaches to manage unique challenges.

Further, as management strives to implement better evidence-based decision making processes through the use of analytics they are confronted with a myriad of possible data types and storage areas which are often inaccessible. Typically data is stored in a variety of separate systems, for example in student information systems, student social media, learning management systems, student library usage, individual computers and administrative systems holding information on programme completion rates and learning pathways. In addition, these data come in different formats of audio, video, text and pictures.

The challenge is to develop meta-systems that can aggregate the various data stores and formats in order to make accessible for more intensive analysis. If achieved, we would gain new insights from this hidden digital activity that currently underpins much of the action associated with teaching, learning and research within higher education. It is only then that we would be in a position to address the complex challenges that institutions face.

This paper contributes to our theoretical understanding of the role Big Data plays in addressing contemporary challenges institutions of higher education face. The paper draws upon emergent literature in Big Data and discusses ways to better utilise the growing data available from various sources within an institution, to help understand the complexity of influences on student-related outcomes, teaching and the ‘what-if questions’ for research experimentation. The paper further presents opportunities and challenges associated with the implementation of Big Data analytics in higher education.

## 2. Emergency of Big Data

The use of data to inform decision-making in organizations is not new; business organizations have been storing and analyzing large volumes of data since the advent of data warehouse systems in the early 1990s. However, the nature of data available to most organizations is changing. According to IBM, 80 percent of the data organizations currently generated are unstructured, and they come in a variety of formats such as text, video, audio, diagrams, images and combinations of any two or more formats. With traditional solutions becoming too expensive to scale or adapt to rapidly evolving conditions, business organizations are looking for affordable technologies that will help them store, process, and query all of their data. Organizations store most of their data that are meant to support decision-making in data warehouses. A data warehouse refers to a central repository of data or a centralized database system used for analyzing and reporting data. Data warehousing also represents an ideal vision of maintaining a central repository data that provides an organization with a living memory of data that can be leveraged for better decision-making.

Recent developments in database technologies made it possible to collect and maintain large and complex amounts of data in many forms and stored in multiple sources within multiple point in time and space. In addition, there are analytical tools available that can turn this complex data into meaningful patterns and value, a phenomenon referred to as Big Data. Big Data describes data that is fundamentally too big and moves too fast exceeding the processing capacity of conventional database systems (Manyika, et., al. 2010).

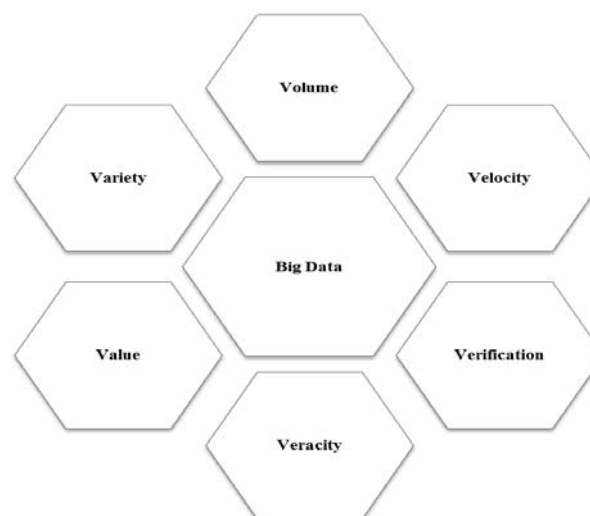
The notion of Big in the term itself is misleading as contested by critics. It does not reflect only data size but complexity. As Yang (2013) pointed out the definition of Big Data has little to do with the data itself, since the analysis of large quantities of data is not new, but rather Big Data includes emergent suit of technologies that can process massive volumes of data of various types at faster speeds than ever before.

## **2.1 Overview of Big Data Research**

Big Data is an emergent knowledge system that is already changing the nature of knowledge and social theory in fields such as business, health and government, while also having the power to transform management decision-making theory. It is a set of techniques, procedures and technologies dealing with voluminous amounts of data in physical or digital formats, data that is being stored in diverse repositories ranging from tangible account bookkeeping records of an educational institution to class test or examination records to alumni records (Sagiroglu & Sinanc, 2013).

The growing interests in Big Data is associated with the sophistication of technologies used to process huge and complex quantities of data and the value accrued to utilization of such data. Douglas (2001) proposed what is commonly known as the “three Vs” (Volume, Velocity and Variety) features of Big Data, but generally, the literature presents a number of fundamental characteristics associated with the notion of Big Data including:

- *Volume*—referring to a large amount of information that is often challenging to store, process, transfer, analyse and present.
- *Velocity*—relating to increasing rate at which information flows within an organization— (e.g. institutions dealing with financial information and relating that to human resources and productivity).
- *Veracity*—refers to the biases, noise and abnormality in data generated from various sources within an institution. It also looks at how data that is being stored, and meaningfully mined to address problem being analysed. Veracity also covers questions of trust and uncertainty associated with the collection, processing and utilization of data.
- *Variety*—referring to data presented in diverse format both structured and unstructured.
- *Verification*—refers to data corroboration and security.
- *Value*—refers to the ability of data in generating useful insights, benefits, and business processes, etc. within an institution.



**Figure 1.** Key characteristics of Big Data

There are also other important properties of Big Data that needs to be taken into account such as data validity, which refers to accuracy of data, and data volatility, a concept associated with the longevity of data and their relevance to the outcomes of the analysis, especially the length of time required to store data in a useful form for further appropriate value-added analysis.

### **3. Analytics and Big Data in Higher Education**

Trillions of data are being collected, and stored in various institutional databases and data warehouses. Many institutions of higher education are increasingly, delivering learning online, subsequently, there is widespread growing availability of online data repositories, educational digital libraries, and their associated tools (Borgman et al., 2008; Choudhury, Hobbs, & Lorie, 2002; Xu & Recker, 2012). Big Data provides innovative techniques and technologies used for capturing, storing, distributing, managing and analysing larger-sized datasets with diverse structures.

In higher education it is concerned with the application of tools and techniques in the analysis of large sets of complex data, with an ultimate goal of improving performance and governance of an institutional system (learning, teaching, and research). This growing area of inquiry incorporates research areas such as educational data mining research which is focused on developing new tools for discovering patterns in educational data and learning analytics a growing area in education an area dedicated largely on examining indicators of individual student and class performance (Luan, 2002; Romero & Ventura, 2010). The utilization of Big Data in higher education involves the interpretation of a wide range of administrative and operational data gathered processes aimed at assessing institutional performance and progress in order to predict future performance, and identify potential issues related to academic programming, research, teaching and learning (Hrabowski III, Suess & Fritz, 2011; Picciano, 2012).

As a developing field within higher education, Big Data is well positioned to address some of the key challenges currently facing institutions of higher education (Siemen, 2011; Siemens & Long, 2011; Dawson, 2013). Wagner and Ice (2012) noted that technological developments served as catalysts for the move toward the growth of analytics in higher education. Capturing and storing data in data warehouse and applying data mining techniques are the foundations for the future activities involved with higher education (Tulasi, 2013). Analytics are attractive approaches in education due to ability of tools and techniques for data processing and analysis. Mayer (2009) noted that the increase in attention to analytics is also driven by advances in computation. For instance smartphones today exceed the computational power of desktop computers, and are more powerful can accomplish tasks that were impossible only a few years ago (Baker, in press).

In addition, there is a growing number of systems intended to leverage Big Data. Systems such as Apache Hadoop, Hortonworks, MapReduce, and Tableau Software are designed to support the use of analytics tools. Further, SAS and IBM SPSS address the substantial challenges of managing data at the scale of the internet (Dean & Ghemawat, 2010; Baker in press). Further, currently the best platforms for harnessing the power of Big Data are open and flexible. They also blend the right technologies, tools, and features to turn data compilation into to data insight.

### **4. Conceptualising Big Data in Higher Education**

In order to understand the value add of Big Data in higher education, we proposed a conceptual framework that provides a foundation for understanding the fundamental components of Big Data in higher education (see Figure 2), and the data sources that contribute to Big Data in higher education institutions environments (figure 3), and the schema needed to harvest and process data into meaningful outputs (figure 4).

#### **4.1 Institutional analytics**

Institutional analytics refers to a variety of operational data that can be analysed to help with effective decisions about making improvements at the institutional level. Institutional analytics include assessment policy analytics, instructional analytics, and structural analytics. They make use of reports, data warehouses and data dashboards that provide an institution with the capability to make timely data-driven decisions across all departments and divisions.



**Figure 2:** Components of Big Data in higher education

#### **4.2 Institutional analytics**

Institutional analytics refers to a variety of operational data that can be analysed to help with effective decisions about making improvements at the institutional level. Institutional analytics include assessment policy analytics, instructional analytics, and structural analytics. They make use of reports, data warehouses and data dashboards that provide an institution with the capability to make timely data-driven decisions across all departments and divisions.

#### **4.3 Information Technology Analytics**

Information technology (IT) analytics covers usage and performance data which helps with monitoring required for developing or deploying technology for institutional use, developing data standards, tools, processes, organizational synergies and policies. Information technology analytics largely aims at integrating data from a variety of systems—student information, learning management, and alumni systems, as well as systems managing learning experiences outside the classroom.

When students interact with learning technologies, they leave behind data trails which can reveal their sentiments, social connections, intentions and goals. Researchers can use such data to examine patterns of student performance over time—from one semester to another or from one year to another, and develop rigorous data modelling and analysis to reveal the obstacles to student access and usability, and to evaluate any attempts at intervention.

#### **4.4 Academic/Program Analytics**

Academic analytics encapsulates all the activities in higher education affecting administration, research, resource allocation and management (Tulasi, 2013). It provides overall information about what is happening in a specific program and how to address performance challenges. Academic analytics combines large data sets with statistical techniques and predictive modelling to improve decision making, and provide data that administrators can use to support the strategic decision-making process as well as provides a method for benchmarking in comparison to other institutions.

#### **4.5 Learning Analytics**

Learning analytics is concerned with the measurement, collection, and analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs (Siemens & Long, 2011). More broadly, references are often made to learning analytics software and techniques are often used to improve processes and workflows, measuring academic and



institutional data and generally improving organizational effectiveness. Although such usage is often referred to as learning analytics, it is more associated with ‘academic analytics’. Learning analytics is undertaken more at the teaching and learning level of an institution and is largely concerned with improving learner success (Jones, 2012).

## 5. Sources and types of Big Data in Higher Education

Big Data within institutions of higher education is made up of various types such as administrative and operational databases. The widespread introduction of learning management systems (LMS) such as Blackboard and Moodle resulted into increasingly large sets of data. Each day, LMS accumulate increasing amounts of students’ interaction data, personal data, systems information and academic information (Romero et al., 2008). LMS keep record of students’ key actions. This include records of logging on to a system, posting and viewing messages, accessing materials, etc. Student data in LMS could also include more detailed information on the content of students’ postings or other writing, their choices and progress through a specific interactive unit or assignment, or their particular preferences and habits as manifested over a range of tasks and interactions or semester (Friesen, 2013).

Big Data also draws sources of data from social media posts, online news articles, digital scans of academic journals, student financial aid profiles, students library usage. Data drawn from social media can also help us understand students’ behaviours. Today, Facebook, Twitter, Instagram and LinkedIn are students’ primary source of information, communication and influence. Figure 2 summarizes possible types of data sources present in many institutions and constitute Big Data.

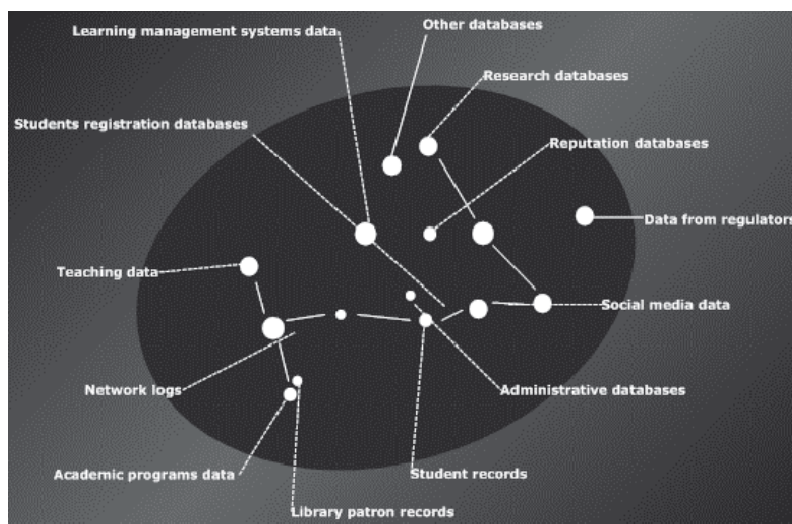


Figure 3: Components of Big Data in higher education

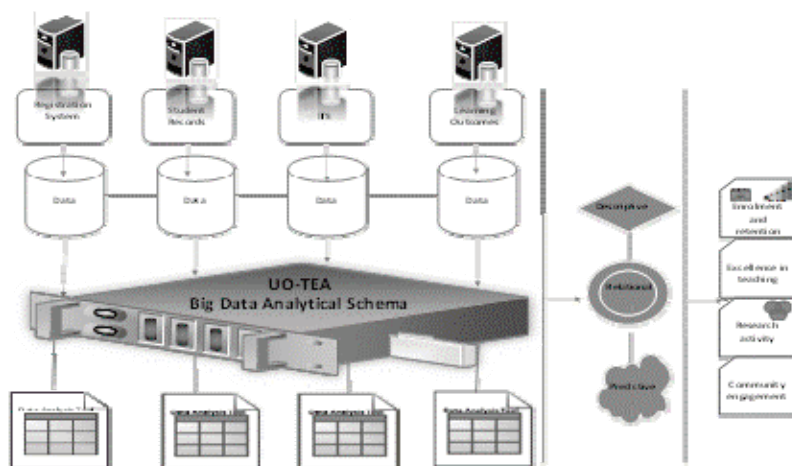


Figure 4: Schema for data processing

## **6. Opportunities**

Big Data in higher education can be transformative, altering the existing processes of administration, teaching, learning and academic work (Baer & Campbell, 2011), contributing to policy and practice outcomes, and helping address contemporary challenges facing higher education (Daniel & Butson, 2013). For instance at the departmental level, Big Data can provide dashboards that reveal patterns confirming program strengths or deficiencies. By designing programs that collect data at every step of the students learning processes, institutions can address student needs with customized modules, assignments, feedback and learning trees in the curriculum that will promote better and richer learning. Similarly, on the institutional level, predictive models can be developed that help intuitions align resources such as tutorials, online discussions, and library assistances, and thus personalizing learning pathways.

The added-value of Big Data in higher education is the ability to identify useful data and turn it into useable information by identifying patterns and deviations from patterns. Schleicher of OECD, 2013 reported that: “Big Data is the foundation on which education can reinvent its business model and build the coalition of governments, businesses, and social entrepreneurs that can bring together the evidence, innovation and resources to make lifelong learning a reality for all. So the next educational superpower might be the one that can combine the hierarchy of institutions with the power of collaborative information flows and social networks.” Siemens (2011) further indicated that “[learning] analytics are a foundational tool for informed change in education” and provide evidence on which to form understanding and make informed (rather than instinctive) decisions.

Furthermore, when used effectively, Big Data can help institutions enhance learning experience and improving student performance across the board, reduce dropout rates and increase graduation numbers. Analytics also provides researchers with opportunities to carry out retrospective analysis of student data, producing predictive models capable of identifying students at risk and providing appropriate intervention (EDUCAUSE, 2011; US Department of Education, 2012). Further, Big Data analytics could be applied to examine student entry on a course assessment, discussion board entries, blog entries, or wiki activity could be recorded, generating thousands of transactions per student per course. This data would be collected in real or near real time as it is transacted and then analyzed to suggest courses of action.

For the individual student dashboards can help them track their own progress and personalize their own learning pathways. Further, the availability of student interaction data can be used to understand students learning patterns online. For instance, by measuring the time a student spends on a module or question, educators can assess how well an assignment performs for any one subgroup of students and identify areas of difficulty for a class or individual learner.

Furthermore, data left in Learning Management Systems (LMS) and database technologies will prepare institutions to proactively plan and stay relevant to their students. For instance, an analysis of a student’s attendance, test results, laboratory performance and class participation (and more), can indicate whether or not he or she will graduate on time. Early intervention can improve retention, and her overall experience. Within a learning management system data obtained could be used to:

- help predict student success in an online course
- monitor a student’s behaviour and engagement level
- notify a lecturer when the student seems to be less engaged with learning
- personalize the learning process
- reduce classroom administrative work
- help lecturer refine content and keep relevant
- monitor students’ progress and provide real-time feedback

## **7. Challenges of implementation**

There are a number of anticipated challenges associated with collection and implementation Big Data in higher education. For instance, the costs associated with collecting, storing, and developing algorithms to mine data can be time consuming and complex. Furthermore, most of institutional data systems are not interoperable, so



aggregating administrative data and classroom and online data for example can pose additional challenges (Daniel & Butson, 2013). While combining data sets from across a variety of unconnected systems can be extremely difficult, it offers better comprehensive insights that inevitably lead to improved capabilities of predictive modelling. Dringus (2012) suggested that one way of overcoming these problems, is to increase institutional transparency by clearly demonstrating the changes that analytics can help to achieve.

Though Big Data also has the potential to help learners and instructors recognize early warn signs (Wagner & Ice, 2012), wide institutional acceptance of analytics requires a clear institutional strategy that help different parties within an institution to work together (Ali et al., 2013). For instance, it requires the involvement of information technology services departments in planning for data collection and use is deemed critical. This observation is consistent with a recent US Department of Education (2013) report suggested that the successful implementation of Big Data in higher institution would depend on collaborative initiatives between various departments in a given institution.

Implementation of Big Data depends on the ability of an institution to co-create data governing structures and delivery of more progressive and better policies and strategies for data utilization and governance (Daniel & Butson; Schleicher, 2013). Wagner and Ice (2012) also pointed out that by increasing collaborative ventures on Big Data initiatives help all groups take ownership of the challenge involving student performance and persistence. Further, Dringus (2012) suggested that the practice of learning analytics should be transparent and flexible to make it accessible to educators (Dringus, 2012; Dyckhoff et al., 2012).

However, in many instances, there is still a divide between those who know how to extract data and what data is available, and those who know what data is required and how it would best be used. Lack of an institutional vision on the added-value of data and inability to leverage collective institutional effort makes collaboration difficult. Further, Romero and Ventura (2010) note, analytics has traditionally been difficult for non-specialists to generate (in meaningful context), to visualize in compelling ways, or to understand, limiting their observability and decreasing institutional impact of analytics (Macfadyen & Dawson, 2011).

Finally, Big Data raises issues on ethics associated with quality of data, privacy, security and data ownership. It also raises the question of an institutions responsibility for taking action on issues based on the information available (Jones, 2012). Dringus (2012) suggests that bringing transparency to learning analytics as a practice could be used to help deter any potentially wrongful use of data.

## **8. Summary and Future directions**

Big Data is being used to convey all sorts of concepts such as huge quantities of data, social media data, advanced databases and algorithms. Whatever the label, business organizations are starting to systematically understand and explore how to process and analyze a vast array of information in new ways that can help them use evidence to improve decision-making. Institutions of higher education are faced with growing challenges that require evidence based decision-making. Large stores of data already exist in most institutions. By analyzing this data, analytics applications have the potential to provide various types of dashboards (institution, program, teacher, and student), which provides predictive view of upcoming challenges. For instance, the ability to mine unstructured and informal connections and information produced by students, including social media, machine sensors and location-based data, will allow educators to uncover useful facts and patterns they were not able to identify in the past. Further, the application of Big Data in higher education opens a window into students' interaction with content, peers, and educators, as well as presenting an opportunity to measure those interactions and draw conclusions about what they mean for student retention and success as well as ability to proactively address policy challenges.

Despite the potentials of Big Data in higher education, there are potential challenges that need to be addressed. This include issues around data management and governance structures associated with Big Data in higher education, as well as maintenance of privacy. As the amount of data available for use is ever-increasing, the benefits will come from good learning management, reliable data warehousing and management, flexible and transparent data mining and extraction, and accurate and responsible reporting.

An institutional research project at the University (Daniel & Butson, 2013) is being undertaken to develop and test conceptual and theoretical underpinnings of Big Data analytics in higher education, as well as developing

key performance indicators, metrics and methods for capturing, processing and visualizing data. In addition, a set of diagnostic tools and an integrated technology enhanced data analytic framework and ultimately a Data warehouse for Big Data Analytics are being pursued.

The research team is also currently engaged in identifying and establishing policies that specifies who is accountable for various portions or aspects of the institutional data and information, including its accuracy, accessibility, consistency, completeness, and maintenance. This research will also be looking at defining processes concerning how data and information are stored, archived, backed up, and protected, as well as developing standards and procedures that define how the data and information are used by authorized personnel and implement a set of audit and control procedures to ensure ongoing compliance with governmental regulations and industrial standards.

## References

- Ali, L., Adasi, M., Gasevic, D., Jovanovic, J., and Hatala, M. (2013). Factors influencing beliefs for adoption of a learning analytics tool: An empirical Study. *Computers & Education*, 62, pp. 130–148.
- Baer, L & Campbell, J. (2011). *Game Changers*. EDUCAUSE
- Base, A. (2013). Five pillars of prescriptive analytics success. *Analytics*, pp. 8-12, March/April Issue [ANALYTICS-MAGAZINE.ORG]
- Baker, R.S.J.d., Inventado, P.S. (in press) Educational Data Mining and Learning Analytics. To appear in J.A. Larusson, B. White (Eds.) *Learning Analytics: From Research to Practice*. Berlin, Germany: Springer.
- Borgman, C. L., Abelson, H., Dirks, L., Johnson, R., Koedinger, K. R., Linn, M. C., Lynch, C. A., Oblinger, D. G., Pea, R. D., Salen, K., Smith, M. S. & Szalay, A. (2008). *Fostering Learning in the Networked World: The Cyberlearning Opportunity and Challenge*. A 21st Century Agenda for the National Science Foundation. Report of the NSF Task Force on Cyberlearning. Office of Cyberinfrastructure and Directorate for Education and Human Resources. National Science Foundation. Retrieved July 12 2014 from [http://www.nsf.gov/publications/pub\_summ.jsp?ods\_key=nsf08204]
- Choudhury, S., Hobbs, B., & Lorie, M. (2002). A framework for evaluating digital library services. *D-Lib Magazine*, 8. Retrieved July 12 2014 from <http://www.dlib.org/dlib/july02/choudhury/07choudhury.html>
- Daniel, B.K. & Butson, R. (2013). Technology Enhanced Analytics (TEA) in Higher Education, *Proceedings of the International Conference on Educational Technologies*, 29 Novemebr-1 December, 2013, Kuala Lumpur, Malaysia, Pp. 89-96.
- Dean, J. & Ghemawat, S. (2010). MapReduce: a flexible data processing tool. *Commun. ACM* 53 (1), pp.72-77.
- Douglas, L (2001). "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner Report. Retrieved 30 December 2013 from [http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf]
- Dringus, L. (2012). Learning Analytics Considered Harmful. *Journal of Asynchronous Learning Networks*, 16 (3).
- Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., and Schroeder, U. (2012). Design and Implementation of a Learning Analytics Toolkit for Teachers. *Educational Technology & Society*, 15 (3), pp. 58–76.
- EDUCAUSE (2011) Learning Initiative, "7 Things You Should Know about First-Generation Learning Analytics," December 2011. Retrieved on July 14 from [http://www.deloitte.com/assets/DcomIreland/Local%20Assets/Documents/Public%20sector/IE\\_PS\\_making%20the%20grade\\_IRL\\_0411\\_WEB.pdf](http://www.deloitte.com/assets/DcomIreland/Local%20Assets/Documents/Public%20sector/IE_PS_making%20the%20grade_IRL_0411_WEB.pdf)
- Friesen, N. (2013). Learning Analytics: Readiness and Rewards. *Canadian Journal of Learning Technology*, 39 (4). Retrieved from <http://www.cjlt.ca/index.php/cjlt/article/view/774>
- Hilbert, Martin, *Big Data for Development: From Information- to Knowledge Societies* (January 15, 2013). Available at SSRN: <http://ssrn.com/abstract=2205145> or <http://dx.doi.org/10.2139/ssrn.2205145>
- Hrabowski, III, F.A., Suess, J. & Fritz, J. (2011). Assessment and Analytics in Institutional Transformation. *Assessment and Analytics in Institutional Transformation (EDUCAUSE Review)*. EDUCAUSE Review, vol. 46, no. 5 (September/October 2011). Retrieved August 8th 2013 from [http://www.educause.edu/ero/article/assessment-and-analytics-institutional-transformation]
- IBM What is Big Data? — Bringing Big Data to the enterprise. 01.ibm.com.
- Jones, S. (2012). Technology Review: the possibilities of learning analytics to improve learner-centered decision-making. *Community College Enterprise*, 18 (1), 89-92.
- Luan, J. (2002). Data mining and its applications in higher education. In A. Serban & J. Luan (Eds.), *Knowledge management: Building a competitive advantage in higher education* (pp. 17-36). Josey-Bass: PA.
- Macfadyen, L. P., and Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers & Education* 54, 588-599.
- Macfadyen, L. P., and Dawson, S. (2012). Numbers Are Not Enough. Why e-Learning Analytics Failed to Inform an Institutional Strategic Plan. *Educational Technology & Society*, 15 (3), 149-163.
- Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute. Retrieved on July 14 2014 from: [http://www.mckinsey.com/Insights/MGI/Research/Technology\\_and\\_Innovation/Big\\_data\\_The\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation).

- Mayer, M (2009). The physics of Big Data. Retrieved on July 14 2014 from <http://www.parc.com/event/936/innovation-atgoogle.html>.
- Picciano, A.G. (2012). The Evolution of Big Data and Learning Analytics in American Higher Education, 16 (3), pp.9-20. Journal of Asynchronous Learning Networks.
- Romero, C. R. & Ventura, S. (2010). "Educational Data Mining: A Review of the State of the Art." IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews 40 (6): 601–618.
- Sagiroglu, S., & Sinanc, D. (2013). Big Data: a review. Proceedings of the international conference on collaboration technologies and systems (CTS), pp. 42–47.
- Schleicher, A. (2013). Big Data and PISA. Retrieved on 4th August 2013 from [\[http://oecdeducationtoday.blogspot.co.nz/2013/07/big-data-and-pisa.html?m=1\]](http://oecdeducationtoday.blogspot.co.nz/2013/07/big-data-and-pisa.html?m=1)
- Schroek, M., Shockley, R., Smart, J., Romero-Morales, D. & Tufano, P. (2012). Analytics: The real-world use of Big Data. How innovative enterprises extract value from uncertain data. Research report: IBM Institute for business value. Retrieved on July 12 2014 from [\[http://www.ibm.com/smarterplanet/global/files/se\\_\\_sv\\_se\\_\\_intelligence\\_\\_Analytics\\_-\\_The\\_real-world\\_use\\_of\\_big\\_data.pdf\]](http://www.ibm.com/smarterplanet/global/files/se__sv_se__intelligence__Analytics_-_The_real-world_use_of_big_data.pdf)
- Siemens, G., and Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE Review, 46 (5), 30.
- Siemens, G. (2011). How data and analytics can improve education, July 2011, Retrieved on August 8 from [\[http://radar.oreilly.com/2011/07/education-data-analytics-learning.html\]](http://radar.oreilly.com/2011/07/education-data-analytics-learning.html)
- Tulasi, B. (2013). Significance of Big Data and Analytics in Higher Education. International Journal of Computer Applications, (pp. 23-25), 68 (14), pp. 23-25
- U.S. Department of Education, Office of Educational Technology, Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief, Washington, D.C., 2012.
- Wagner, E. & Ice, P. (2012). Data changes everything: delivering on the promise of learning analytics in higher education. Educause Review JULY/AUGUST 2012, pp. 33-42.
- Xu, B., & Recker, M. (2012). Teaching Analytics: A Clustering and Triangulation Study of Digital Library User Data. Educational Technology & Society, 15 (3), 103–115.
- Yang, L. (2013). Big Data Analytics: What is the Big Deal? Retrieved July 12 2014 from [\[http://knowledge.ckgsb.edu.cn/2013/12/30/technology/big-data-analytics-whats-big-deal/\]](http://knowledge.ckgsb.edu.cn/2013/12/30/technology/big-data-analytics-whats-big-deal/)

# Data, Information and Intelligence

Scott Erickson<sup>1</sup> and Helen Rothberg<sup>2</sup>

<sup>1</sup>Ithaca College, Ithaca, USA

<sup>2</sup>Marist College, Poughkeepsie, USA

[gerickson@ithaca.edu](mailto:gerickson@ithaca.edu)

[hrothberg@aol.com](mailto:hrothberg@aol.com)

**Abstract:** As big data and business analytics continue to grow and draw attention, there is also an increasing recognition that existing theory and conceptual development in other areas studying intangible assets may have something of value to add. The authors continue a research stream exploring the connections between knowledge management, competitive intelligence, and big data/business intelligence. This includes theory development, comparing the concepts of the different fields and looking at where contrasting emphases can add value through cross-fertilization of ideas. The stream also includes comparison of methods and techniques, from big data platforms to knowledge management (information technology solutions, communities of practice, etc.) and on to competitive intelligence analysis tools (e.g. environmental scanning, war games). While further developing themes from some earlier work, such as the role of business analytics in recognizing the value of basic data and information and the similar contribution of knowledge management to encouraging and capturing insights from intangible assets, this paper will look more specifically at the potential contribution of competitive intelligence to our understanding of all these fields. Data are available on the industry level concerning big data capabilities and knowledge management/intangible asset development. To these are added further data, specifically on competitive intelligence activity and threats in comparable industries. Focusing on competitive intelligence (CI) can bring new insights to the conversation. CI has always valued the full range of intangible asset inputs (data, information, and knowledge) and actionable intelligence, something knowledge management can neglect (with its strict definitions of valuable knowledge vs. mere data or information). CI can also be more directed, looking for additional data, information, or knowledge in a specific area in order to address a specific question. This paper will look at data on competitive intelligence activity in specific industries, identifying those with high intelligence commitment as opposed to those without. These results will be compared and contrasted with data on big data potential and significant development of intangible assets, also by industry. As a consequence, the authors are able to prescribe directions for the development of all, some, or none of the disciplines in question while also providing recommendations for cross-field combinations for greater impact.

**Keywords:** Knowledge management, intellectual capital, data, information, big data, business analytics, competitive intelligence

---

## 1. Knowledge

Intangible assets and their role in obtaining and maintaining competitive advantage have a considerable history in the economics and business literatures, albeit in different forms and with different definitions. But combining the different literature streams into a coherent whole can help us understand the opportunities presented by attention to these intangibles as well as a better grasp of the potential of the relatively new interest in big data and business analytics.

The earliest interest in intangibles generally revolved around innovation. Given that context, it's not surprising that Schumpeter (1934) is often seen as an important source, especially his focus on knowledge combination, learning, and new insights. Nelson & Winter's (1982) evolutionary theory suggested that intangibles such as skills and learning might be the real drivers of competitive advantage in firms. This view fit nicely with the near-contemporary development of the resource-based view of the firm (Wernerfelt 1984), with organizational knowledge as the key resource or differentiator. From there, a knowledge-based view of the firm, with knowledge and similar intangibles at the heart of competitive advantage was a natural extension (Teece 1998; Grant 1996). Indeed, many in the field now argue that organizational knowledge may be the only source of unique, sustainable competitive advantage as other, more traditional assets and differentiators become commoditized.

As may be discerned from the previous paragraph, it's very easy in this area to throw terms around loosely, with somewhat ambiguous concepts like learning, knowledge, know-how, and skills being used in precise applications. Fortunately, most of the disciplines have settled on definitions following Ackoff's (1989) DIKW hierarchy, positing that intangibles grow from basic data to ordered information, providing opportunities for users to obtain knowledge which can then pass to wisdom. In more current applications, intelligence, or actionable knowledge/information/data, is often used in place of wisdom at the highest level. These

definitions recur in specific disciplines and have some importance in setting the boundaries and areas of study for fields. In knowledge management (KM), for example, data are simply observations, information is data in context, and knowledge itself is information subjected to experience and reflection (Zack 1999b). In this field, the data and information inputs are of interest as sources, but only the knowledge and insights created in employees' heads are really worth creating systems to manage. In this view, data and information have little value on their own. And intelligence is left to other fields altogether. This leaves some gaps related to the more recent interest in big data business analytics, though other disciplines pay more attention. As we'll see, these differing perspectives create opportunities for cross-fertilization between disparate disciplines.

As just alluded to, knowledge management (KM) as well as the related discipline of intellectual capital (IC) focus on the value of knowledge assets. IC is more about identifying, cataloguing, and measuring knowledge assets (Bontis 1999; Edvinsson & Malone 1997; Stewart 1997). The field has generally settled on the areas of human capital (individual knowledge), structural capital (organizational routines, culture, etc.), and relational capital (knowledge about external relationships), though we'll see some extensions shortly. KM, on the other hand, focuses on what to do with these intangible assets. In doing so, scholars and practitioners have explored differences in the nature of knowledge that can make it more of a challenge to develop. In particular, the differences between tacit and explicit knowledge (Nonaka & Takeuchi 1996; Polanyi 1967) have been highlighted, as well as characteristics such as complexity and stickiness (McEvily & Chakravarthy 2002; Zander & Kogut 1995; Kogut & Zander 1992). The nature of the organizations looking to manage knowledge have also been studied, focusing on absorptive capacity (Cohen & Levinthal 1990), social capital (Nahapiet & Ghoshal 1998), social networks (Liebowitz 2005) and related topics.

With these types of circumstances in mind, KM practitioners can move to better managing their particular knowledge assets. KM is all about more effectively employing and growing knowledge by combination, sharing, socialization, and other such methods (Zack 1999a; Grant 1996). The appropriate tools for growth can be subject to the variables just mentioned, so theory and practice both emphasize assessing and reacting by using tacit tools for tacit knowledge, explicit for explicit knowledge, etc. (Choi & Lee 2003; Schulz & Jobe 2001; Boisot 1995). From this perspective, approaches ranging from communities of practice to IT solutions have been developed and employed (Brown & Duguid 1991; Matson, Patiath & Shavers 2003; Thomas, Kellogg & Erickson 2001).

But, as noted at the beginning of the section, KM and IC both focus largely on knowledge assets. Precursors like data and information or extensions like intelligence are rarely discussed. But that doesn't mean there isn't an emphasis elsewhere, in related fields.

## **2. Data, Information, Knowledge, and Intelligence**

Over the past decade, new views of intangible assets have resulted in new directions in strategic decision-making. Big data and business analytics are only the latest step, albeit a very important one. In one prominent example, Andreou, Green & Stankosky (2007) looked to make sense of a variety of related approaches by creating the "List of Operational Knowledge Assets" including market capital (competitive intelligence, enterprise intelligence), human capital, decision effectiveness, organizational capital, and innovation and customer capital. While many could be included in a standard human, structural, relational capital taxonomy, LOKA does make explicit newer directions such as competitive intelligence and enterprise intelligence that don't fit neatly in that structure. A broader view of intangible assets also begs the question of whether data, information, and intelligence have more of a role to play in competitive advantage than is commonly thought in the traditional KM/IC discipline.

One direction for exploring these questions is to look specifically at competitive intelligence (CI). CI has a more extensive history than some of the other "intelligences", growing in both practice and in academia over the past few decades. Basically, IC is about understanding and anticipating competitor actions by identifying, collecting, and analyzing relevant data, information, and knowledge (Prescott & Miller 2001; Gilad & Herring 1996; Fuld 1994). The result is actionable intelligence. CI has a number of distinctions from KM/IC as well as the readily apparent similarities (in particular, both focus on the value of intangible assets (Rothberg & Erickson 2005; 2002). Initially, CI tends to have a wider range of intangible inputs, including the data and information aspects already mentioned several times. Practice also combines these disparate inputs with specialized analysis techniques and applications (Fleisher & Bensoussan 2002; McGonable & Vella 2002). CI is

similar to KM/IC in that operations mature with added sources of inputs and increased analysis experience of the coordinating team (Wright, Picton & Callow 2002; Raouch & Santi 2001). Prior to and at the analysis stage, however, CI is much more likely to conduct targeted searches to fill specific information gaps and, of course, is also designed for action (Gilad 2003; Bernhardt 1993). KM/CI certainly encourage action but the focus of activities is often more about building the knowledge base and the network for sharing it.

This contrast between KM/IC and CI provides a natural entry to a deeper discussion on big data and business analytics. A number of these topics should seem very familiar to anyone with experience with the latter. Big data, by definition, is about massive amounts of data and information, the wider view of what intangible assets are valuable, and the impact of big data comes through business analytics processes able to organize, analyze, and find higher level insights from these databases. Indeed, the big data, business analytics, business intelligence area often references prior work in KM and these other areas (Bose 2009; Jourdan, Rainer & Marshall 2008).

For ease of reference, we'll refer to the whole area as either big data or business analytics from here on out. The seeds of current excitement about business analytics comes from rapid increases in the power of modern IT systems. Substantially decreased costs for storage and processing, including in the cloud, have enabled organizations to save more and more data while doing more and more analysis of the resulting massive databases (Bussey 2011; Vance 2011b). Much of the data or information collected revolves around operations, supply chain, and channel performance; transactional and customer information, and communications (including social media) (Vance 2011a). According to a much-cited McKinsey Global Services report, big data add value with greater transparency and more immediate feedback on performance, an ability to experiment in real time, provide opportunities for more precise segmentation, rationalized decision-making, and generating new product ideas (Manyika, et. al. 2011).

Scholarly development in this new field is limited. One does see repeated mention of the "three V's" (and sometimes additional V's), referencing data volume, velocity, and variety (Laney 2001). All have increased with the drop in IT costs, allowing the increased storage and higher-level analysis already mentioned, potentially leading to better decision-making at all levels (Beyer & Laney 2012). Metrics to date have generally centered on data storage (Manyika, et. al. 2011) and case studies are used to good effect in explaining the details (Liebowitz 2013). As we know, however, it's not just the size of the databases that is important, it's what's done with them. Per KM/IC, CI, and related disciplines, the really valuable part of data and information comes from the higher level insights, the knowledge and intelligence derived from their analysis. A human element, through analytical techniques, is necessary to the process (Zhao 2013).

And this is an important area where previous research in KM/IC and related fields may have something to offer. KM perspectives would certainly be enhanced by attention to other intangible assets beyond knowledge as well as the added capabilities found in analysis and actionable intelligence. But the KM side does have its own contributions as there is deep scholarship and practice concerning growth in knowledge assets from combination and insight as well as how to engage humans to interact with IT systems (Matson, Patiath & Shavers 2003; Thomas, Kellogg & Erickson 2001). Person-to-person and person-to-system issues such as perceived usefulness, motivation, trust, and other such matters have been at the forefront of the field for over twenty years.

Given our interest in finding connections between existing scholarship in KM/IC and CI, a comparison of how each field sees the world may be interesting. In previous work, we've looked at connections between KM/IC and big data (Erickson & Rothberg 2014). In this study, we combine data specifically on competitive intelligence practice with that on big data capabilities.

### **3. Big Data and Competitive Intelligence**

In order to provide a framework for discussion, we created Table 1 from two sources. Initially, there is information concerning big data, taken from a McKinsey Global Institute (MGI) report (Manyika, et. al. 2011). This is combined with industry categorizations based on levels of intangible assets and competitive intelligence activity (Erickson & Rothberg 2013; 2012). From this table, we can begin to suggest some ideas concerning the



relationship between big data and knowledge as well as what underlying concepts may explain differences present in the information.

**Table 1** Big Data, Knowledge, and Competitive Intelligence, by Industry

Industry	Stored Data per Firm (terabytes)	Stored Data, US Industry (petabytes)	SIC	Competitive Intelligence Activity	Firm/CI Activity
Security & investment Services	3,866	429	6211	13	8.74
Banking	1,931	619	6020	23	14.27
Communications & Media	1,792	715	48	25	16.34
Utilities	1,507	194	49	19	6.94
Government	1,312	848			
Discrete Manufacturing	967	966	36,37,38	40,29,45	8.97
Insurance	870	243	63	82	3.49
Process Manufacturing	831	694	2834	78	10.96
Resource Industries	825	116	13, 24	12,3	9.60
Transportation	801	227	40-45	7	41.46
Retail	697	364	53-59	23	23.25
Wholesale	536	202	50-51	27	14.29
Health Care Providers	370	434	80	4	300.28
Education	319	269	82	5	172.70
Professional Service	278	411	731	6	252.32
Construction	231	51	16	2	113.04
Consumer & Recreational Services	150	105	79	4	179.20

The first three columns are taken straight from the MGI report, including the industry definitions, though sorted according to Stored Data per Firm for our purposes. Stored Data by US Industry was sourced by MGI from research firm IDC and is an estimate of the total data held by firms with more than 1,000 employees in each broadly defined industry. This number is then divided by number of firms to get the per firm figure in the second column. Per firm obviously provides a much different assessment as the number of firms varies dramatically between concentrated industries like those in financial services and dispersed industries such as manufacturing. Figures are from 2008.

The latter three columns come from our database via a Fuld & Company survey. Standard Industrial Classification (SIC) was used to sort the data by industry, breaking firms with similar products and operations into appropriate groups for comparison purposes. Sometimes this included SIC level one classification, sometimes this included groupings down to the second, third, and fourth levels. Only groupings with at least 20 observations (firm and year, financial returns, at least \$1 billion in annual revenue) were included in the database. The SIC classifications were matched as closely as possible with the more general descriptions of the MGI report.

Data from the Fuld & Company database includes self-reports on the maturity and professionalism of each CI operation. These were weighted by level (4 for highest degree of proficiency, 1 for lowest) and combined with number of firms reporting each to develop the index shown in the CI Activity column. Typically, we use the total amount of CI activity within an industry or industry sector as a proxy for how aggressive the intelligence-gathering environment might be. Here, however, given that the industries and SIC categories between the two databases don't necessarily match up well, and that the number of firms in the different industries vary so dramatically, we thought it best to also look at some per firm assessment of CI activity. This is presented in the final column (number of firms calculated from the two "stored data" columns and then divided by the CI metric. This figure ends up being the inverse of CI activity per firm, so a higher value indicates less relative CI emphasis.

#### 4. Discussion

The data are very revealing. On just the level of CI activity per industry sector, there is some evidence of more activity where higher levels of big data are present, with double-digit CI activity scores. Essentially, there are more, and more professional, CI operations in data-heavy industries. But these metrics are also less than consistent, with the highest levels of activity closer to the middle big data. There are likely explanations, as we'll see, but these initial results are also a reason to look more carefully at the data.

Based on the additional adjustment, calculating the per firm CI metric to go with the per firm big data metric makes sense. The number of firms in the extremely broad MGI industry classifications don't necessarily match up well with the number in the SIC categories for IC. Indeed, if one closely examines the data, some industries with relatively high numbers of CI activity also have an abnormally large number of firms represented (retail, wholesale). Similarly, some of those with seemingly muted CI numbers have a relatively low number of resident firms (securities, utilities, natural resources). Consequently, the per firm CI score not only provides more information but is probably the truer representation.

And it does provide some very interesting results. Just as the stored data per firm column shows a clear progression from top to bottom, so does the CI activity per firm column. The level of CI activity for industries with a commitment to big data varies within a fairly small range but is of a magnitude difference from the level for industries with low commitment. Scores in the single digits and low double digits (for CI) go along with a high degree of data storage per firm. Scores in the triple digits go along with a low degree of data storage per firm. The pattern is clear and convincing. High levels of CI activity appear to go along with high levels of investment in big data capabilities. Big data has the potential to feed competitors' CI operations, and those pursuing big data strategies should probably be sure to invest in data protection, counterintelligence, and their own CI operations.

But there is even further depth to the data, given what we know about KM/IC and CI in some of these industries. Those with the absolute lowest CI ratios: insurance, utilities, and securities are all regulated industries generating a lot of operational and transactional data. Not all of that data is particularly interesting (movements of money, investments, power). Indeed, what we know from KM/IC is that these industries do not rate real highly for intangible assets (Erickson & Rothberg 2012). What is valuable is the rare tacit insight, the eureka moments that result in new directions in such industries (new portfolio ideas, new lending or investment strategies, whatever is new in today's old-line utilities). In such environments, the ability of CI operations to spot new ideas at competitors, based on changes in big data patterns or discovering the insight itself can result in rapid copying. Consequently, CI can be both effective and profitable by cutting down periods of new product or new process exclusivity.

Similarly, the next lowest CI ratios are in manufacturing industries, both process (pharmaceuticals, plastics, chemicals) and discrete (machinery, transportation). Again, there is value in CI, not necessarily from the supply chain or operations data but from what it tells observers is going on in terms of R&D and new products, process optimization, or other manufacturing improvements. Further, just as in the previous group, actual activities tend to be hidden from view, so CI operations of a certain maturity are needed to peek behind the veil.

At the other end of the spectrum, a very different pattern emerges. Those with very high scores on the metric (indicating low levels of CI activity) are invariably services. Services can employ big data, as they have operations, transactions, and communications just like other industries. Services, however, have unique characteristics that make them hard to manage, including intangibility, perishability, producer variability, and customer involvement. Consequently, optimizing and standardizing processes can be difficult, leaving a gap in terms of what big data is able to accomplish. Moreover, the services seen here are often right out in the open, making advanced CI operations something amounting to overkill. Much of what can be learned could come simply from walking through a public facility and observing the layout or operation. In such situations, less worries exist about CI, so there is less need for protection (or, as pointed out, important matters are so transparent as to be almost impossible to protect). There also appears to be less need for counterintelligence or heavy investment in a CI operation.

Overall, there appears to be a link between a high level of data storage, indicative of industries and firms investing in a big data approach, and competitive intelligence interest in those big data stores. Even though the direct relationship between the two metrics is clear, there are also less obvious insights in the results. The type of data harvested in big data approaches is different, depending on industry, and the insights or knowledge gained from such data also differ. The initial results show great promise, but there is additional potential apparent if further and deeper analysis is done.



## 5. Conclusions

Numerous fields look at intangible assets as a valuable corporate resource, ranging from traditional innovation studies to knowledge management, intellectual capital, and competitive intelligence. The newer entity of big data and business analytics has drawn great interest for its potential in helping firms craft contemporary competitive strategies and tactics. Big data and business analytics hold tremendous promise, but these newer disciplines do not exist in a vacuum. There is a strong likelihood that similarities exist with the other fields and that cross-fertilization is possible, helping us to understand all disciplines better as we move forward.

This paper looks specifically at competitive intelligence activities, at the intelligence/wisdom end of the DIKW hierarchy, opposite big data's data/information end. Results drawn from a study of big data by McKinsey and our own database on CI activity show a strong relationship between the fields. Higher investment in big data capabilities appears to go along with higher levels of competitive intelligence activity. Competitors would appear to be interested in big data holdings (and business analytics results). Further analysis of the relationship suggests that there are differences in the nature of the big data and related knowledge insights that can lead to further distinctions in the pattern of big data and competitive intelligence activity.

## Acknowledgement

The authors gratefully acknowledge the contributions of Fuld & Company for providing some of the data used in this study.

## References

- Andreou, A.N., Green, A. & Stankosky, M. (2007) "A Framework of Intangible Valuation Areas and Antecedents", *Journal of Intellectual Capital*, 8(1), 52-75.
- Bernhardt, D. (1993). *Perfectly Legal Competitive Intelligence—How To Get It, Use It and Profit from It*, Pitman Publishing, London.
- Beyer, M.A. & Laney, D. (2012) "The Importance of 'Big Data': A Definition", retrieved from <https://www.gartner.com/doc/2057415>.
- Boisot, M. (1995) "Is Your Firm a Creative Destroyer? Competitive Learning and Knowledge Flows in the Technological Strategies of Firms", *Research Policy*, 24, 489-506.
- Bontis, N. (1999) "Managing Organizational Knowledge by Diagnosing Intellectual Capital: Framing and Advancing the State of the Field", *International Journal of Technology Management*, 18(5-8), 433-462.
- Bose, R. (2009) "Advanced Analytics: Opportunities and Challenges", *Industrial Management & Data Systems*, 109(2), 155 – 172.
- Brown, J.S. & Duguid, P. (1991) "Organizational Learning and Communities-of-Practice: Toward a Unified View of Working, Learning, and Innovation", *Organizational Science*, 2(1), 40-57.
- Bussey, J. (2011) "Seeking Safety in Clouds", *The Wall Street Journal*, (September 16), B8.
- Choi, B. & Lee, H. (2003) "An Empirical Investigation of KM Styles and Their Effect on Corporate Performance". *Information & Management*, 40, 403-417.
- Cohen, W.M. & Levinthal, D.A. (1990) "Absorptive Capacity: A New Perspective on Learning and Innovation", *Administrative Science Quarterly*, 35(1), 128-152.
- Drucker, P.F. (1991) "The New Productivity Challenge", *Harvard Business Review*, 69(6), 69-76
- Edvinsson, L. & Malone, M. (1997) *Intellectual Capital*, Harper Business, New York.
- Erickson, G.S. & Rothberg, H.N. (2013, "Competitors, Intelligence, and Big Data", in Liebowitz, J., ed., *Big Data and Business Analytics*, CRC Press, Boca Raton, FL., 103-116.
- Erickson, G.S. & Rothberg, H.N. (2012) *Intelligence in Action: Strategically Managing Knowledge Assets*, Palgrave Macmillan, London.
- Fleisher, C.S. & Bensoussan, B. (2002) *Strategic and Competitive Analysis: Methods and Techniques for Analysing Business Competition*, Prentice Hall: Upper Saddle River, NJ.
- Fuld, L.M. (1994) *The New Competitor Intelligence: The Complete Resource for Finding, Analyzing, and Using Information About Your Competitors*, John Wiley, New York.
- Gilad, B. (2003) *Early Warning: Using Competitive Intelligence to Anticipate Market Shifts, Control Risk, and Create Powerful Strategies*, ANACOM, New York.
- Gilad, B. & Herring, J., eds. (1996) *The Art and Science of Business Intelligence*, JAI Press, Greenwich, CT.
- Grant, R.M. (1996) "Toward a Knowledge-Based Theory of the Firm", *Strategic Management Journal*, 17(Winter), 109-122.
- Jourdan, Z., Rainer, R.K. & Marshall, T.E. (2008), "Business Intelligence: An Analysis of the Literature", *Information Systems Management*, 25(2), 121-131.
- Kogut, B. & Zander, U. (1992) "Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology", *Organization Science*, 3(3), 383-397.

- Laney, D. (2001) "3D Data Management: Controlling Data Volume, Velocity and Variety", retrieved November 1, 2013 from <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.
- Liebowitz, J. ed. (2013) *Big Data and Business Analytics*, CRC Press/Taylor & Francis: Boca Raton, FL.
- Liebowitz, J. (2005). Linking social network analysis with the analytical hierarchy process for knowledge mapping in organizations. *Journal of Knowledge Management*, 9(1), 76-86.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Hung Byers, A. (2011) *Big Data: The Next Frontier for Innovation, Competition and Productivity*, McKinsey Global Institute.
- Matson, E, Patiath, P. & Shavers, T. (2003) "Stimulating Knowledge Sharing: Strengthening Your Organizations' Internal Knowledge Market", *Organizational Dynamics*, 32(3), 275-285.
- McEvily, S. & Chakravathy, B. (2002) "The Persistence of Knowledge-Based Advantage: An Empirical Test for Product Performance and Technological Knowledge", *Strategic Management Journal*, 23(4), 285-305.
- McGonagle, J. & Vella, C. (2002) *Bottom Line Competitive Intelligence*, Quorum Books, Inc., Westport, CT.
- Nahapiet, J. & Ghoshal, S. (1998) "Social Capital, Intellectual Capital, and the Organizational Advantage", *Academy of Management Review*, 23(2), 242-266.
- Nelson, R.R. & Winter, S.G. (1982) *An Evolutionary Theory of Economic Change*, Harvard University Press, Cambridge, MA.
- Nonaka, I. & Takeuchi, H. (1995) *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press, New York.
- Polanyi, M. (1967) *The Tacit Dimension*, Doubleday, New York.
- Prescott, J.E. & Miller, S.H. (2001) *Proven Strategies in Competitive Intelligence: Lessons from the Trenches*, John Wiley and Sons, New York.
- Raouch, D. & Santi, P. (2001) "Competitive Intelligence Adds Value: Five Intelligence Attitudes", *European Management Journal*, 19(5), 552-559.
- Rothberg, H.N. & Erickson, G.S. (2005) *From Knowledge to Intelligence: Creating Competitive Advantage in the Next Economy*, Elsevier Butterworth-Heinemann, Woburn, MA.
- Rothberg, H.N. & Erickson, G.S. (2002) "Competitive Capital: A Fourth Pillar of Intellectual Capital?", in Bontis, N., ed., *World Congress on Intellectual Capital Readings*, Elsevier Butterworth-Heinemann, Woburn, MA.
- Schulz, M. & Jobe, L.A. (2001) "Codification and Tacitness as Knowledge Management Strategies: An Empirical Exploration", *Journal of High Technology Management Research*, 12, 139-165.
- Schumpeter, J.A. (1934) *The Theory of Economic Development*, Harvard University Press, Cambridge, MA.
- Stewart, T.A. (1997) *Intellectual Capital: The New Wealth of Nations*, Doubleday, New York.
- Teece, D.J. (1998) "Capturing Value From Knowledge Assets: The New Economy, Markets for Know-How, and Intangible Assets", *California Management Review*, 40(3), 55-79.
- Thomas, J.C., Kellogg, W.A. & Erickson, T. (2001), "The Knowledge Management Puzzle: Human and Social Factors in Knowledge Management", *IBM Systems Journal*, 40(4), 863-884.
- Vance, A. (2011a) "The Data Knows", *Bloomberg Businessweek*, (September 12), 70-74.
- Vance, A. (2011b) "The Power of the Cloud", *Bloomberg Businessweek*, (March 7), 52-59.
- Wernerfelt, B. (1984) "The Resource-Based View of the Firm", *Strategic Management Journal*, 5(2), 171-180.
- Wright, S., Picton, D. & Callow, J. (2002) "Competitive Intelligence in UK Firms, A Typology", *Marketing Intelligence and Planning*, 20(6), 349-360.
- Zack, M.H. (1999) "Developing a Knowledge Strategy" *California Management Review*. 41(3), 125-145.
- Zack, M.H. (1999b) "Managing Codified Knowledge", *Sloan Management Review*, 40(4), 45-58.
- Zander, U. & Kogut, B. (1995) "Knowledge and the Speed of Transfer and Imitation of Organizational Capabilities: An Empirical Test", *Organization Science*, 6(1), 76-92.
- Zhao, D. (2013) "Frontiers of Big Data Business Analytics: Patterns and Cases in Online Marketing", in J. Liebowitz, ed., *Big Data and Business Analytics*, CRC Press/Taylor & Francis: Boca Raton, FL.

# Building an Undergraduate Program in Business Analytics

Scott Erickson, William Tastle and Danielle Puleo

Ithaca College, Ithaca, USA

[gerickson@ithaca.edu](mailto:gerickson@ithaca.edu)

[tastle@ithaca.edu](mailto:tastle@ithaca.edu)

[dpuleo1@ithaca.edu](mailto:dpuleo1@ithaca.edu)

**Abstract:** This paper reviews some of the thought process toward establishing a business analytics program at a small undergraduate institution. The case for big data and business analytics as a field of study is reviewed, especially concerning its growth as a business application and as a source of jobs in the upcoming decade. Illustrative programs, both graduate and a few of the limited number of undergraduate in analytics are benchmarked and show a pattern of content in statistics, data applications, and functional analytics. A proposed track/concentration is included, with current courses and proposed courses, some of which will require additional resources. Overall, such an approach is an opportunity to establish a unique academic offering providing exceptional value to students but also illustrates the creativity necessary to establish such a program with limited resources.

**Keywords:** Business analytics, business intelligence, curriculum, undergraduate

---

## 1. Introduction

The growth of interest in business analytics/business intelligence in graduate programs has surged over the last few years. With increasing demand for students who understand analytics, some growth in undergraduate programs has already begun, albeit quite modestly. In anticipation, this paper reports on the experiences of one small institution engaged in developing an undergraduate business analytics program.

Data on the characteristics of the few programs already established (at all levels) sheds light on the typical requirements for study of business analytics. The process is then a blending of skills required of students and topical coverage with available faculty resources, while keeping an eye towards the continuing evolution of the discipline and needs of business. Given the size of this institution, special challenges exist that may prove interesting to those looking at similar programs at all levels (or looking to hire from similar programs).

This institution, an AACSB-accredited business school, has recently taken steps toward encouraging students to complement basic technology knowledge, predominantly spreadsheets, with more advanced skills leading to official certification. Such a background will enable students in the program to move into solution-specific coursework in programming, statistics, artificial intelligence, data mining, and business intelligence. Additional analytical/interpretive capabilities would be provided in an application course and strategy capstone.

The particular challenges come from existing resources. A business school with only 600 majors will have limited staff in some functional areas, such as information systems. In this case, a minimal staff consisting of one tenured faculty member augmented by one or two adjunct or non-tenure-eligible instructors exists in information systems. As a result, the importance of separating out the basic technology training, as well as an external quality standard, is critical. The tenure track faculty member needs to deliver the core solution-specific coursework, so the program requires flexibility in providing the requisite preparation for that advanced work.

Similarly, an ability to utilize other resources for the remaining aspects of the program is also important. Close external relationships with corporate partners providing business intelligence tools are extremely useful in stretching resources. Similarly, the ability to bring in other disciplines, such as marketing, finance, and accounting, also expands the pool of instructors possessing the requisite background knowledge as well as adding tools and methods suitable for analytics (e.g. SAS, R). In particular, an existing marketing analytics course and another tenured faculty member with a background in knowledge management and competitive intelligence provides an opportunity to lend perspective to the program, guiding students in how analytics are used in specific business applications. Opportunities are also available in areas like operations and finance.

The advantages of the approach are a program that punches above its weight in terms of apparent capabilities and a chance to establish a unique, differentiated offering at the undergraduate level. The disadvantages are

the obvious reliance (even over-reliance) on specific faculty members and external relationships. All points and supporting data are developed in more detail and with appropriate discussion in the full paper and presentation.

## **2. Business Analytics**

The past few years have seen a burst of interest in big data, business analytics, business intelligence, and related areas. Although specific definitions of the terminology can differ and all of the subject areas behind the terms have both similarities and dissimilarities, most of the conversation centers on a few basic concepts. These concepts relate to applying different approaches and techniques to the study of large amounts of data and information. One consequence has been a growth of interest in training students in these approaches and techniques, particularly at the graduate level. As the growth continues, there are likely to be attractive opportunities for graduating students at the undergraduate level as well. Studies routinely predict significant and well-paid job growth in business analytics fields (e.g. Waller 2014; Burtch 2013).

The explosion of interest in big data is a direct result of advancements in data collection, storage, and processing capabilities. Huge amounts of data are regularly generated by supply chain and enterprise systems, transactions systems (cash registers to web sales), and communications (Google Analytics, social media). With drops in the cost of storing and analyzing such data, opportunities have been created to find insights that can contribute to better strategic and tactical decisions (Vance 2011a). This process of mining the big data for useful insights is referred to in various applications as business intelligence, business analytics, marketing intelligence, marketing analytics, and other such terms (Bussey, 2011; Vance, 2011b). From such analyses flow capabilities such as real-time experimentation, tighter segmentation, more objective algorithm-driven decision-making, and wider and deeper idea generation for innovation (Manyika, et. al., 2011), any of which can lead to competitive advantage and superior marketplace performance. Similarly, a report completed for SAS suggests significant potential in cost reduction, time reduction, new offerings based on big data, and more support for internal business decisions (Davenport & Dyché 2013).

From practice, theory has also begun to develop. A good amount of the academic work has centered around the “three V’s” (or more, there is a cottage industry in adding new “V’s” (DeSouza 2014)): data volume, input and output velocity, and data variety (Laney 2001). Low-cost computing, especially in the cloud, has allowed substantial growth in all three areas, and new analysis techniques have accompanied the growth (Beyer & Laney, 2012). Metrics have started to develop (Manyika, et. al., 2011) as have case histories of applications (Liebowitz, 2013). This base of computing capacity and availability are critical to big data and the reason the buzz word has developed. And it is important to develop students able to build the systems that generate, harvest, and organize the data. But it’s also critical to remember that it means little without some contribution to better decision-making. So there is also an analysis and intelligence component requiring some degree of human judgment. Without the analysis and decision-making skills, business analytics remains just a big database (Zhao 2013). Indeed, the Manyika, et. al. (2011) study already cited suggests that the ability to take advantage of big databases is dependent on four factors: talent, data-driven mindset, IT intensity, and data availability. The first two deal almost entirely with human skills and attitudes and the last is about human/systems interactions. To succeed in this new world, students will need a variety of training and abilities.

As a consequence of all these trends, the demand for business school graduates who possess expertise in business analytics and related fields is growing at an explosive rate. The Bureau of Labor Statistics doesn’t list a business intelligence analyst domain, but peripheral domains abound with demand projections, including market research analysts, operations research analysts, and statisticians. Projections suggest a significant skills gap and talent shortage, with not enough prepared individuals entering the field (Davenport & Dyché 2013; Brown, Court & Willmott 2013). Unfortunately, business schools are generally not prepared to address this important and emerging field, especially at the undergraduate level.

Table 1 includes information from some illustrative undergraduate approaches to business analytics and is followed by a table with some illustrative graduate programs, chiefly drawn from our geographic region. In drawing examples for the table, we specifically differentiated between data analysis programs typically housed in everything from Arts & Sciences or Engineering to Science & Technology or Business as opposed to business analytics programs housed exclusively in Schools of Business. The former were often focused on building

databases and statistical analysis without necessarily providing business applications or training in decision support. Instead, we report the more fully formed business-specific programs.

**Table 1** Illustrative Undergraduate Programs

<b>Institution</b>	<b>Curriculum</b>
Babson College	<b><u>Business Analytics Concentration</u></b> Business Intelligence & Data Analytics Cases in Business Analytics (3 Electives in Finance, Marketing, MIS, or Quantitative Methods)
McIntire School of Commerce University of Virginia	<b><u>Business Analytics Track</u></b> Database Management Systems Marketing Research Techniques Business Analytics (1) Elective in Marketing, Finance, or Knowledge Management
Fisher College of Business Ohio State University	<b><u>BS Data Analytics (proposed)</u></b> (3) Math courses (5) Statistics courses (9) Computer/Information Science courses <b><u>Business Analytics Specialization</u></b> Business Analytics: Principles & Concepts Business Analytics: Applications & Experience (3 Elective courses in Finance, Accounting MIS, Customer Insights, or Operations & Logistics)

**Table 2** Illustrative Graduate Programs

<b>Institution</b>	<b>Curriculum</b>
LeBow College of Business Drexel University	<b><u>MS Business Analytics</u></b> MIS (3 courses) Operations Research (2 courses) Statistics (3 courses) Business Analytics capstone 3 courses in MIS, OR, Stats, or functional area
Lally School of Management Rensselaer Polytechnic Institute	<b><u>MS Business Analytics</u></b> Business Core (Stats & Operations, Intro to Acctg and Finl Mgmt, MIS, 3 courses) Analytics Core (3 required, 2 elective courses) Domain-Focused electives (2 courses)
Graduate School of Business Bentley College	<b><u>MS Business Analytics</u></b> Foundation: Stats (1 course) Analytics Core (6 courses, Comp Science, Math, Stats) Electives (4 courses, functional areas)
School of Business Quinnipiac University	<b><u>MS Business Analytics</u></b> CIS (3 courses) Business Analytics (4 courses, including capstone) Electives (4 courses, chiefly operations)
Leonard N. Stern School of Business New York University	<b><u>MS Business Analytics</u></b> Module 1 (Digital Mktg, Stats, Data Mining, Decision Making) Module 2 (Network Analytics, Decision Models) Module 3 (Operations, Advanced Decision Models, Data Visualization) Module 4 (Strategy, Markets) Module 5 (Capstone)

As shown by these examples, business analytics curricula are often characterized by three areas of study: database foundations, math and statistics, and decision-making, often in a functional area. Degrees tending more toward data analysis are often more focused on the first two areas. Those with more of a business emphasis, often with more business coursework, will include more of the latter two areas. As noted throughout the examples, actual application choices run the gamut of the business disciplines, from finance to marketing, from operations/logistics to accounting. The less fully developed undergraduate approaches are both scarcer (not nearly as many programs available) and more limited in scope, not providing the depth and breadth of the graduate experiences.

### **3. Background and Resources**

In the case of our institution, we represent a college with an overall enrollment of 6,200. The School of Business includes roughly 700 majors, supported externally by a separate Math Department and a separate Computer Science Department, both housed in the School of Humanities and Sciences. The School of Business includes less than 30 full-time-equivalent faculty. These resources support undergraduate degrees in business administration, accounting, and legal studies and a small MBA program.

The program's interest in a business analytics program is grounded in a few key faculty members, some serendipitous resources and connections, and an ability to design the program to fit the circumstances. As suggested by the benchmarking above, a full-scale business analytics program will include a database competency, a statistical analysis competency, and an intelligence/analytical competency (often framed by a specific discipline). In our case, the statistical component is largely left to external departments, with specific courses delivered by the Math Department (Statistics for Business), Psychology Department (Statistics for Psychology), and the Communications School (Audience Research). The Business School is left to deliver the database and analysis components.

The former could conceivably be aided by a cooperative arrangement with the Computer Science (CS) Department, but a curricular change just two years ago dropped the CS course required in the business curriculum given its uneven coverage and delivery. Some aspects of the course many students already knew or could easily gain the necessary knowledge through workshops offered by Information Technology Services. More critical aspects of the course weren't always effectively delivered. As a result, it was easier, in some ways, to just take over delivery of those aspects (chiefly Excel basics). Though some re-staffing of Computer Sciences with new, full-time faculty may allow other collaborations in the future.

The information technology/management information systems area in the Business School includes only a single tenured professor. This faculty member is responsible for delivery of the core business curriculum course in Business Systems & Technology (BST). With changes over time, this course now chiefly delivers the spreadsheet content and practice we believe contemporary business need, including Microsoft Excel certification for all students. An additional higher level spreadsheet elective has been offered in some semesters and an experimental lead-in course is currently being piloted (allowing the BST course to start at a higher level and include both advanced spreadsheet topics and an introduction to database software). But a single faculty member can't teach all these alternatives. There has been some support, particularly for the experimental course, through adjuncts.

Other faculty in the School of Business teach discipline-specific courses. Marketing Research was altered to become Marketing Analytics a few years ago, based on input from the school's Business Advisory Council, observation of industry trends, and conversations with recent alumni. We'll discuss some of the specifics shortly, but the basic shift was away from survey design and administration and toward handling communication and observation databases, perhaps collected by survey but often not. Few other existing courses show much promise for a business analytics program though faculty backgrounds in areas such as Radio Frequency Identification tags (Operations Management) and Knowledge Management/Competitive Intelligence (Marketing) do offer potential value.

### **4. Structuring a Program**

Currently our students are exposed to an introduction to business intelligence only in the BST course, but that is too little to give students any more than a cursory introduction to the domain. Hence, we seek to construct a program whose successful completion would prepare our students for internships and employment. Our



rough outline for a suitable undergraduate program follows, including coursework already in the general business core, additional courses already available or feasible with current resources, and some stretch courses requiring commitment of additional resources.

### **Mathematics**

*Statistics for Business, Economics and Management* (existing) is a course already required for all Business Administration and Accounting majors. Taught out of the Math Department in the School of Humanities & Sciences, this is a freshman or sophomore level course establishing the fundamentals of probability and statistics. Applications are emphasized, especially in some alternative versions of the course (*Statistics for Psychology*), and SPSS is also taught in some sections. Without this foundation, the study of analytics is either impossible or without substance.

### **Database Construction and Management**

*Technology and Applications* (existing, experimental) is a freshman level course specifically designed to bring all students up to a standard level of computer-based competence. It is not unusual to hear faculty talking about the sophistication of new freshmen, but in fact there is virtually no depth to any level of computer knowledge. While these new students have been “nursed” by computers and know every new piece of social network programs, once one moves off the surface there is a void of knowledge. Even with respect to the “simple” word processing program once one moves beyond the typing of words and perhaps the insertion of a graphic, sophisticated knowledge is nonexistent. Word processing, presentation software, spreadsheets and database knowledge at a level such that one can be productive at non-trivial tasks should be the norm for entry into any other computer-based courses. Students should have the opportunity to become Microsoft Core certified in two of the Office products.

*Business Systems and Technology (BST)* (existing) is a sophomore level course required of all Business and Accounting majors. The course has evolved in recent years, and will evolve further if the experimental course above prepares students more fully. BST takes students from the advanced novice level and moves them to the advanced level of using the spreadsheet as a two dimensional modeling tool. This course includes some basic training in R and an introduction to business intelligence. The successful student will be Microsoft Excel Expert certified. Students should know how to program functions, build array functions, and understand how to nest functions to accomplish very complex operations. At this level students are expected to study the text themselves to become certified while classroom instruction is on the application and usage of the more complicated functions.

*Computer Programming I* (proposed) would also a freshman or sophomore level course. Programming knowledge is essential to creating non-trivial functions in spreadsheets or other programs, develop efficient macros, allows for the application of programming skills in programs such as R, and most importantly, permits the student to learn how to think sequentially. This is arguably the best course (except possibly mathematics) for developing thinking and problem solving skills and so should be available to anyone, even outside the business analytics track.

*Data Management in Business* (proposed) would be a junior level course encompassing database concepts through SQL, creating competent users of the R statistical program, and developing knowledge of artificial intelligence software used in data mining. Much of this course would be devoted to the application of these tools to the mastery of data mining techniques.

*Principles of Business Intelligence* (proposed) would be a junior level course building a thorough understanding of descriptive, predictive and prescriptive analytics, and developing expertise in using industrial-grade business intelligence software.

### **Functional Applications**

*Marketing Analytics* (existing) is an introduction to analysis of data related to marketing decisions, including program data in marketing information systems and project data collected through primary research. Includes both analysis of existing databases and the most common and practical problems associated with collecting and analyzing new data. Emphasis is on both qualitative methods and quantitative methods. Computer assignments require students to apply latest software packages (students have a choice of SAS or SPSS).

As mentioned earlier, the Marketing Analytics course makes sense to include in the program. The shift to analytics from research was intentional, as trends in the field have dramatically changed the nature and amount of data available to marketing decision-makers. Traditional marketing research courses heavily feature survey research, often including the development and administration of a questionnaire. Marketing as a discipline, however, is using more commercial data (Nielsen, Google Analytics), more observation data (loyalty programs, e-commerce, social media), and, even when surveys are used, often more longitudinal data with true panels, measuring the same items and respondents over time.

Our conclusion from these trends was that our students were more and more likely to be consumers of data rather than producers. Consequently, they needed to understand how and from where the data came, but surveys would only be a piece of that. Much more important would be manipulation and analysis of the data, wherever it came from, and an ability to make better marketing decisions. From this kind of preparation, even those going into marketing analytics would be better prepared than if we only taught them survey design. They can hire people to create surveys. The value-added positions would be in interpretation and decision-making. And this perspective once again squares with trends found in big data, as observers stress the importance of illustrative applications (IBM 2013) to gain acceptance and use by decision makers as well as the development of frontline tools that make the databases accessible and usable (Brown, Court & Willmott 2013).

Ideally, the curriculum would expand to include elective choices for students in the area. Students interested in marketing could take the marketing analytics course, those interested in other functional areas would have other options. Marketing is the obvious initial place to offer a functional course, already including an analytics experience, and the discipline also has faculty with training at the SAS campus in North Carolina. But, as noted, possibilities with other faculty in the operations area (experience at the Wal-Mart RFID Center at the University of Arkansas) and in finance will be pursued.

*Strategic Analytics* (proposed) would be the capstone experience. In structuring a capstone, we have an opportunity to take advantage of a faculty member's deep background in knowledge management, competitive intelligence, and related areas. While not obviously related to a business analytics program, there are connections. Ackoff (1989) first discussed some key definitions with his DIKW pyramid, suggesting a hierarchy starting with raw data, organized into information, enhanced with know-how and learning to become knowledge, and then applied as wisdom. If one substitutes intelligence for wisdom, we still talk about the same areas today.

Consequently, data and information contained in the databases providing the foundation for business analytics are considered precursors to the deeper understanding found in knowledge and intelligence. Indeed, big data and business intelligence are often seen as the natural next step following directly from knowledge management and related fields (Bose, 2009; Jourdan, Rainer & Marshall, 2008). Wider views beyond knowledge management, venturing into areas like marketing intelligence, competitive intelligence, and such also clearly bring advanced databases and their analysis into the discussion (Andreou, Green & Stankosky 2007). In short, there are considerable interrelationships and the considerable literature and theory that have developed related to knowledge management, competitive intelligence and such fields can provide important insights for business analytics students. The nature of data, information, and knowledge, the best tools for developing and leveraging them, human interactions with IT systems for managing these intangibles, and basic human considerations such as motivation (to use the systems), trust, social capital and organizational culture, social networks, and so forth offer a valuable high-level view of the entire field. They also offer a unique perspective on the field that can be a differentiator for the program and for our students.

*Internship.* If an internship in data analytics/business intelligence can be identified, students should be encouraged to participate at least in the junior summer. By that time they can perform at an acceptable level.

## **5. Conclusions**

The burgeoning demand for graduates with demonstrated expertise in data analytics and business intelligence has recently been at the forefront of academic conferences (Information Systems Education Conference, Association for Computer Information Systems, Business Intelligence Innovation Summit, Gartner BI & Analytics Summit, and many, many others). The need to offer business intelligence and data analytics is being pushed upon us. To satisfy these growing needs, projected to extend well into the next decade, our institution



has taken steps to construct some sort of program within the business administration degree, allowing students to extend their knowledge of BI and analytics and provide potential recruiters with a transcript that officially denotes their skills and expertise.

A BI program in an institution such as ours must build on the more demanding disciplines of statistics, programming, and Excel but only at the most advanced levels. Once these fundamental tools have been mastered at a sophisticated level of operation, additional mastery of business intelligence can ensue using an industry grade tools such as PowerOLAP (Paris Technologies). Students must be competent in the concepts, terminology, design and construction of user-friendly dashboards and reports to service all the different demands of a business from the executive level through the operational levels and down to the customer first-contact individual level. Students should be able to competently apply not only spreadsheet technology to a solution, they must know when the circumstances warrant a different tool such as R, when confronted with business problems in which, through their standard business education, they fully understand the problem, can address solutions desired by an individual, and can go beyond the request and be proactive in providing a solution that might have been unknown to the requester.

Beyond these tools, students must also be able to ask and then understand the answers to the right questions. Knowing where the data come from and the strengths and weaknesses of the data collection techniques is also important. Understanding how to motivate and incentivize the humans interacting with the data systems so that data, information, and knowledge are contributed to the system (and then made use of) is also something that matters. Also important are analytical techniques for processing these intangible assets in order to make strategic decisions (e.g. scenario planning, war games). The most qualified students will emerge from the program with the statistical and database construction/manipulation skills allowing them to build and use big data. They will also have the analytical talent to process the results, draw insights, and make positive decisions for their organizations.

## References

- Ackoff, R. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, 16, 3–9.
- Andreou, A.N., Green, A. & Stankosky, M. (2007) "A Framework of Intangible Valuation Areas and Antecedents", *Journal of Intellectual Capital*, 8(1), 52-75.
- Bontis, N. (1999) "Managing Organizational Knowledge by Diagnosing Intellectual Capital: Framing and Advancing the State of the Field", *International Journal of Technology Management*, 18(5-8), 433-462.
- Bose, R. (2009) "Advanced Analytics: Opportunities and Challenges", *Industrial Management & Data Systems*, 109(2), 155 – 172.
- Brown, B., Court, D. & Willmott, P. (2013) "Mobilizing your C-Suite for Big Data Analytics", *McKinsey Quarterly*, (November), 1-11.
- Burtch, L. (2013) *The Burtch Works Study: Salaries for Big Data Professionals*, [http://www.burtchworks.com/register\\_study.php](http://www.burtchworks.com/register_study.php), accessed 3/28/2014.
- Bussey, J. (2011) "Seeking Safety in Clouds", *The Wall Street Journal*, (September 16), B8.
- Davenport, T. & Dyche, J. (2013) *Big Data in Big Companies*, SAS Institute (May).
- DeSouza, K.C. (2014) *Realizing the Promise of Big Data: Implementing Big Data Projects*, IBM Center for Business in Government, [www.businessofgovernment.org](http://www.businessofgovernment.org).
- Edvinsson, L. & Malone, M. (1997) *Intellectual Capital*, Harper Business, New York.
- Erickson, G.S. & Rothberg, H.N. (2013) "Competitors, Intelligence, and Big Data", in Liebowitz, J., ed., *Big Data and Business Analytics*, CRC Press, Boca Raton, FL., 103-116.
- Erickson, G.S. & Rothberg, H.N. (2012) *Intelligence in Action: Strategically Managing Knowledge Assets*, Palgrave Macmillan, London.
- Fleisher, C.S. & Bensoussan, B. (2002) *Strategic and Competitive Analysis: Methods and Techniques for Analysing Business Competition*, Prentice Hall: Upper Saddle River, NJ.
- Fuld, L.M. (1994) *The New Competitor Intelligence: The Complete Resource for Finding, Analyzing, and Using Information About Your Competitors*, John Wiley, New York.
- Gilad, B. (2003) *Early Warning: Using Competitive Intelligence to Anticipate Market Shifts, Control Risk, and Create Powerful Strategies*, ANACOM, New York.
- Gilad, B. & Herring, J., eds. (1996) *The Art and Science of Business Intelligence*, JAI Press, Greenwich, CT.
- IBM Center for the Business of Government (2013) *From Data to Decisions III: Lessons from Early Analytics Programs*, (November), [www.businessofgovernment.org](http://www.businessofgovernment.org).
- Jourdan, Z., Rainer, R.K. & Marshall, T.E. (2008), "Business Intelligence: An Analysis of the Literature", *Information Systems Management*, 25(2), 121-131.

- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Hung Byers, A. (2011) *Big Data: The Next Frontier for Innovation, Competition and Productivity*, McKinsey Global Institute.
- Matson, E, Patiath, P. & Shavers, T. (2003) "Stimulating Knowledge Sharing: Strengthening Your Organizations' Internal Knowledge Market", *Organizational Dynamics*, 32(3), 275-285.
- Rothberg, H.N. & Erickson, G.S. (2005) *From Knowledge to Intelligence: Creating Competitive Advantage in the Next Economy*, Elsevier Butterworth-Heinemann, Woburn, MA.
- Stewart, T.A. (1997) *Intellectual Capital: The New Wealth of Nations*, Doubleday, New York.
- Thomas, J.C., Kellogg, W.A. & Erickson, T. (2001), "The Knowledge Management Puzzle: Human and Social Factors in Knowledge Management", *IBM Systems Journal*, 40(4), 863-884.
- Vance, A. (2011a) "The Data Knows", *Bloomberg Businessweek*, (September 12), 70-74.
- Vance, A. (2011b) "The Power of the Cloud", *Bloomberg Businessweek*, (March 7), 52-59.
- Waller, N. (2014) "Big Data: Love It or Face 'Permanent Pink Slip'," *The Wall Street Journal*, (April 9), B6.

# Using Entropy for Healthcare Analytics and Risk Management in Influenza Vaccination Programs

Eugene Levner<sup>1</sup>, Sharon Hovav<sup>2,3</sup>, Hanan Tell<sup>1,2</sup>, Alexander Ptuskin<sup>4</sup>, Avi Herbon<sup>2</sup>

<sup>1</sup>Ashkelon Academic College, Ashkelon, Israel

<sup>2</sup>Bar-Ilan University, Ramat Gan, Israel

<sup>3</sup>Clalit Health Services, Tel Aviv, Israel

<sup>4</sup>Moscow Bauman Technological University (Kaluga Branch), Kaluga, Russia

[eleverner@acad.ash-college.ac.il](mailto:eleverner@acad.ash-college.ac.il)

[sharonh@clalit.org.il](mailto:sharonh@clalit.org.il)

[hanan.tell@biu.ac.il](mailto:hanan.tell@biu.ac.il)

[aptuskin@mail.ru](mailto:aptuskin@mail.ru)

[avher@bezeqint.net](mailto:avher@bezeqint.net)

**Abstract:** Annual influenza epidemics impose great losses in both human and financial terms. A key question arising in large-scale vaccination programs is to balance program costs and public benefits. For a vaccination supply chain (SC) consisting of one or several manufacturers, distribution centers, warehouses, pharmacies, clinics, and customers, we seek to decrease the total monetary expenses of all stakeholders while taking into account public benefits at a nation-wide level. Risks occur in the SC due to a stochastic nature of the vaccination process which fluctuates from year to year, depending on many factors that are difficult to predict and control: adverse events like the supply delay and mismatch between the vaccine stock and demand. It is impractical and unnecessary to construct the entire network of the vaccination supply chain which may have tens of thousands components. Instead, the entropy-based analytics permits to essentially reduce the SC model without discarding essential prognostic information about the possible risks. Entropy is a measure of the uncertainty in a random environment. Extending classical Shannon's entropy concept used in information theory, we use the term to quantify and evaluate the expected value of the information contained in a SC with uncertain but predictable data about the costs and benefits. Knowing the history of adverse events, we estimate the entropy and knowledge about the risks occurring in the vaccination SC, evaluate the loss, define most vulnerable components in the SC and reduce its size. An integer programming model is proposed in which the problem of minimizing the total loss is effectively solved on a reduced SC. This new analytics approach permits to balance the manufacturing, inventory and distribution costs with public benefits and to reduce the incurred losses. A case study is used to test the suggested methodology - we consider the nation-wide vaccination program carried out by the CLALIT Health Services (Israel). The entropy approach permits us to essentially decrease the model size in practice: whereas the initial supply chain in the CLALIT which is constructed for real-life data for year 2013 contains about 1,200 clinics located in different geographical places and serving about one million people grouped into 6,000 population clusters, the reduced supply chain contains only 24 clinic clusters and 120 population sub-groups. Then the mathematical programming problem for minimizing the total vaccination program costs has been formulated on the reduced SC and applied retroactively for the 2013 data. The problem has been solved by the GAMS software and permitted to decrease the annual total expenses by 12%.

**Keywords:** Vaccination, influenza, supply chain, risk, entropy, loss minimization

## 1. Introduction

Influenza is a highly contagious airborne viral infection. Seasonal influenza epidemics cause up to 500,000 deaths per year worldwide, and impose great economic losses as a result of healthcare costs and lost productivity (Chick et al. 2008). Billions of dollars are spent annually for influenza vaccination in an attempt to avoid even greater losses. The current work proposes an analytics driven research approach for enhancing the efficiency of influenza vaccination programs, using supply chain (SC) concepts and techniques. We seek to minimize the total cost of the vaccination supply chain (VSC) while upholding the individual interests of its stakeholders, with their risks being taking into account.

The influenza VSC includes vaccine manufacturers, healthcare organizations (HCOs) (consisting of distribution centers (DCs), pharmacies, clinics, and hospitals), and end users. The latter is the population of individuals to be vaccinated. It is segmented into sub-populations, each characterized by an age range. The cost of the vaccination program includes direct and indirect medical costs of the vaccination process, and direct and indirect non-medical costs of the vaccination supply chain (these terms will be discussed below). The costs of the vaccination program must be balanced against its public benefits. These benefits include, among other

factors, reductions in the number of patient visits, in the number of hospitalization days among different groups of people, in the number of working days lost, and in mortality.

Our study is based on the following premises and assumptions.

- We seek to balance the costs and benefits of the vaccination program, from the perspective of the healthcare organization (HCO), which is a key component between the manufacturer and end users.
- The hierarchical healthcare supply chain can be presented as a tree-type graph.
- The DC's role, as part of the HCO, is reflected in the general objective function by the DC's direct non-medical costs, i.e., inventory holding costs.
- Clinics are responsible for the direct medical costs related to the service of administering vaccine injections to consumers and for the corresponding inventory costs.

Given weekly (or monthly) forecasts of customer demand and limited capacities of the manufacturer, DCs and clinics, the objective is to minimize the costs of the vaccination program while taking into account the program's main public benefits and population risks. We model the problem and present the case study based on data from the CLALIT Health Services, a leading HCO in Israel's national influenza vaccination supply chain. This case study indicates that the integrated approach we propose can lead to substantial savings (up to millions of dollars annually). One of the key planning problems in managing a vaccination supply chain is to coordinate and integrate individual interests of all the players within entire supply chain, i.e., the geographically distributed SC components.

The budget for decision-making aimed at reducing the risks is limited, and therefore the processing of the complete volume of information for all the components of the supply chain can be prohibitively time- and resource-consuming. In spite of the importance of the latter issue, the majority of published analytical works did not handle this, assuming that the size of the SC is known in advance and fixed. A motive for the present research, distinguishing it from other works is: How to reduce a huge size of an original supply chain model so as not to discard meaningful predictive information and to identify those SC components that are the major causes of loss in the chain. This is the key question, to answer which we use Shannon's information entropy (Shannon, 1951) as a measure of our knowledge about risks and their impact. The main idea is that there is no need to elaborate a more and more detailed SC model if a huge model does not lead to an essential change in entropy value, and, hence, in our knowledge.

The remainder of the paper is structured as follows. A review of previous work is given in the next section. A problem description and a corresponding graph model of the hierarchical SC are given in section 3. In section 4, we present an entropy-based algorithm for reducing the problem's size. Then in section 5 we present the mathematical model for optimizing the min-cost vaccination process. A case study in section 6 aims to demonstrate how the model works in practice. We conclude with a summary and directions for future research in section 7.

## **2. Previous work**

In recent decades, researchers have used a variety of analytical approaches to study service quality in healthcare supply chains, the prevailing being the optimization techniques, heuristics and meta-heuristics, and simulation-based analytics. Our study belongs to the first group. Gerdil (2003) presents the main steps of the vaccine production process. This study focuses on product design and manufacturing as a complex cyclical process but does not explore crucial subsequent steps in the vaccination process over the entire supply chain. In contrast, in the present study we explicitly introduce additional stakeholders into the supply chain, namely, the DCs, hospitals pharmacies, clinics and population groups.

A paper by Chick et al. (2008) studies an influenza supply chain and focuses on a specific issue of designing a contract between a manufacturer and a government. This interaction is modeled as a game. They show that a global social optimum can be achieved by using the contract to share production risks between the manufacturer and the government, and consider in depth various game-theoretic aspects of the vaccination SC management.

Recall some basic definitions. An *event* is the observable discrete change in the state of a VSC or its components. An *adverse event* is undesirable unforeseen event such as disruptions, breakdowns, defects, mistakes in the design and planning, shortages of material in the HSC, etc. A *risk driver* is a factor, or a driving force, that may be a cause of the adverse event, We study situations where we can register the adverse events in the considered VSC, during a pre-specified period of time. The registration list is called a *risk protocol*, which provides information about adverse events, their risk drivers and possible loss. Statistics accumulated in the risk protocols permits us to quantitatively evaluate contribution of each driver and the total entropy-based observable information in the SC. The idea is that the prognostic information about the risk probabilities can be inferred from the patterns (events) of past observations and can be used to forecast the future probabilities.

As many other authors, we accept that the *risk* is the expected value of an undesirable outcome, that is, the product of two entities: the *probability* of an undesirable event and the *impact* or *severity* (that is, an expected loss in the case of the disruption affecting the performance of a supply network (Levner & Proth (2005), Kogan & Tapiero (2007). Due to increasing dynamics and growing uncertainty in the healthcare SC environment, risk management became a key concern. In order to reduce the risk level in the supply chain, we need to process data about failures and faults, their causes and economic consequences, locations and occurrence frequencies.

There exists a wide diversity of risk types in the SC. Their taxonomy lies beyond the scope of this paper; we refer to Olson (2012). We will look at the Shannon's information entropy as a tool for measuring the whole information about the risks and potential loss. Many works used the entropy for evaluating the *SC complexity* (see, e.g., Allesina et al., 2010; and Isik, 2010).

Durowoju et al. (2012) and Herbon et al. (2012) explored the fact that a high level of entropy (or chaos) in the supply chain has the effect of impeding perfect SC performance by building obstacles to its supply and delivery; the bigger the obstacle, the more uncertain the state of the entire SC, and as a consequence, a larger amount of information is required to monitor and manage the system. From this perspective, the entropy is regarded as a measure of estimating the knowledge about risks. Durowoju et al. (2012) and Herbon et al. (2012) considered simplest linear supply chains containing only four components - the retailer, distributor, manufacturer and supplier, and did not handle the specificity of risk management in multi-layer supply chains *with precedence constraints* between the SC nodes. This limitation is overcome in the present paper.

The present study continues and extends the latter works by taking a network structure and precedence relations into account. We suggest an iterative method that computes the entropy for the nodes belonging to the same layer of the hierarchical SC successively, layer by layer. The main aim is to reduce the SC size and pick out the most risky components in the reduced SC model. The method respects the precedence relations between the nodes and reduces the SC size without the loss of essential information about the risks. Then the min-cost problem is solved on the reduced SC.

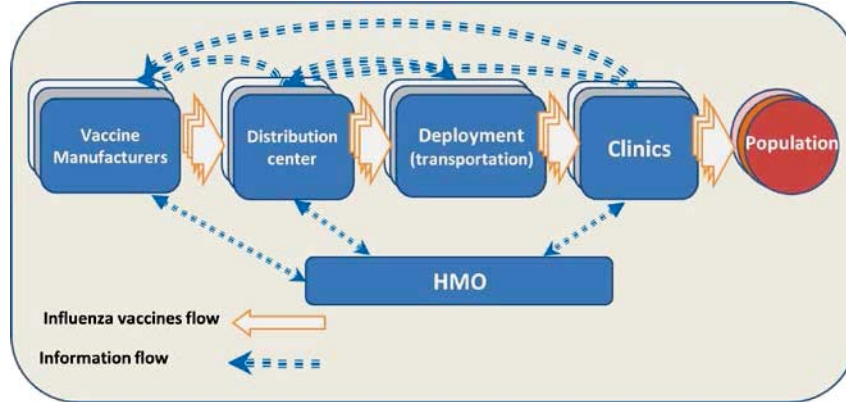
Minimizing the expenses is a critical element in achieving healthcare supply chain effectiveness. The SC literature offers various models and solution methods for solving the min-cost problems. The efficiency and effectiveness of influenza vaccination have been well documented by Nichol (2008). Colombo et al. (2006) uses the cost-benefit analysis of the influenza vaccination comparing the universal mass vaccination policy and the targeted vaccine program, a policy that focuses on high-risk groups only. Our research also addresses this issue, from the SC managerial point of view. The corresponding literature is vast, and we refer to the recent excellent reviews by Tang & Musa (2010), Olson (2012), and Matinrada et al. (2013) that provide the summary of the most commonly used optimization models and methods in SC published in recent years.

### **3. Problem description**

In our analysis, in line with studies in the field of analytical medical management, we divide vaccination costs into four categories: direct and indirect medical costs and direct and indirect non-medical costs. Direct medical costs include product costs, costs of service from nurses and physicians (calculated as average service time multiplied by the average salary of nurses and physicians). Indirect medical costs consist of administration and organizational costs. Direct non-medical costs are logistical costs incurred by the DC, hospitals and clinics. Indirect non-medical costs are organizational/administrative costs by the DC, hospitals and clinics.

Although the papers discussed above provide valuable insights into the costs and benefits of vaccination programs, they do not provide sufficient quantitative analysis that can be used for optimization purposes. Our paper is an attempt to bridge this gap, focusing on direct medical and direct non-medical costs in the framework of a cost-benefit analysis.

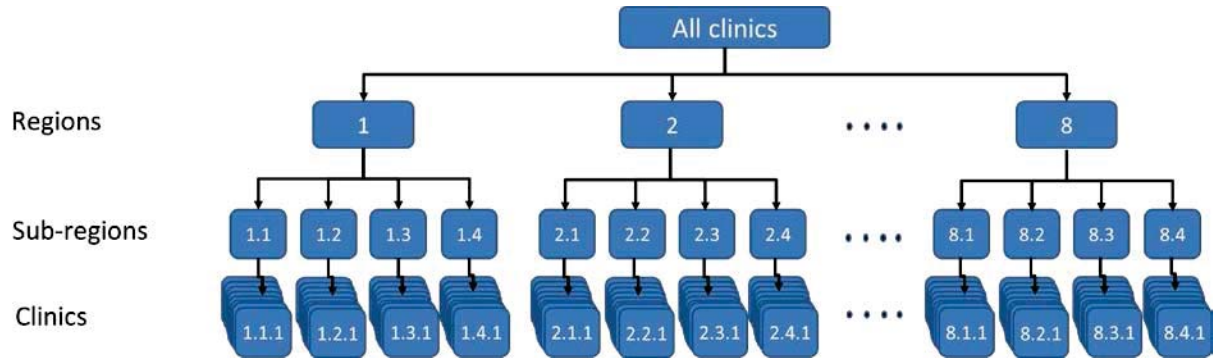
Our supply chain model includes multiple manufacturers and multiple customers. The influenza vaccination SC is illustrated in Fig. 1.



**Figure 1.** A typical vaccination supply network

The vaccination supply chain has a hierarchical structure, and therefore, can be presented as a tree-like graph, in which the upper node ("a root node") depicts the entire VSC without indentifying its elements, the first tier of nodes contains its main components, that is, manufacturers, distribution centers, clinics and population. Each subsequent tier presents the same VSC in a more detailed form. As an illustration, Fig. 2 shows a fragment of a SC in which all the clinics are depicted hierarchically.

The problem considered in the paper is twofold: first, to reduce the model size without discarding essential prognostic information about possible losses, and, second, to minimize the total costs for the vaccination on the reduced SC model before and within the epidemic season, subject to population's demands in vaccines.

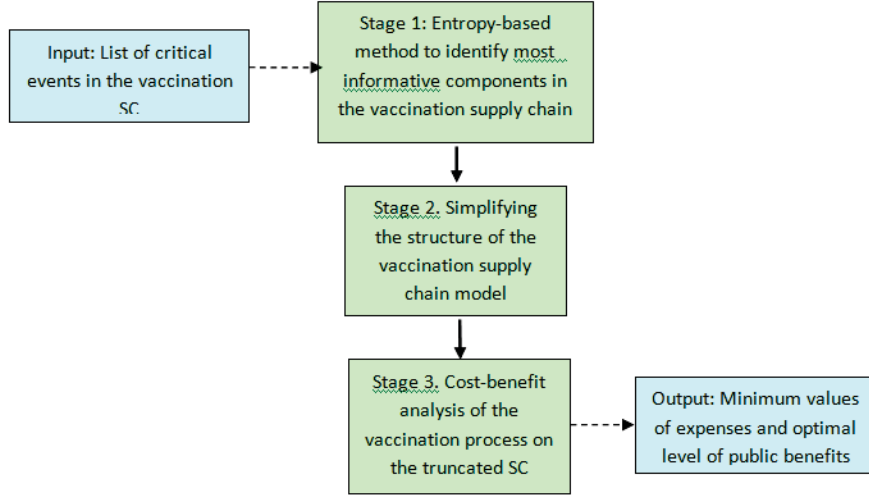


**Figure 2.** Hierarchical structure of a typical vaccination supply network

In the next section we consider the sub-problem of identifying the most significant risk factors and most vulnerable components in a VSC using the entropy approach. As a result, the set of selected VSC components becomes more compact and more controllable than the entire SC. In Section 5, we study the optimization problem aimed at reducing the total costs in the VSC. The problem becomes more compact, controllable and easier for solving due to the decrease in the problem size.



A modeling process elaborated in this paper consists of three basic stages presented in Fig. 3.



**Figure 3.** The logic of the modeling process

#### 4. Entropic approach to reducing the SC model size

Basing on the statistical information about adverse events in the SC, we wish to forecast which risk types in the supply chains that we can expect in the near future are most hazardous. The volume of such information can be prohibitively large. The storage and analysis of all the information on the operation of the entire SC becomes a complex and often just untreatable problem. A natural way to treat such a problem is to cut down the volume of the available information without losing any essential information. Our first goal is to identify those of the SC components which are major sources of information about the risks and losses in the chain. In recent years, *the information entropy* became a widespread measure of the knowledge about where the most important sources of risks in the supply chain are concentrated (see Allesina et al., 2010; Isik, 2010; Herbon et al. 2012, and the references therein).

We start with the definition of Shannon's entropy. In [information theory](#), *entropy* is a measure of the uncertainty, the chaos, the absence of knowledge. It is commonly accepted that the entropy taken with the opposite sign is equivalent to the [information content](#), or the knowledge (Shannon, 1951; Isik, 2010; Stone, 2014).

We have a volume of statistical data, namely, registered information about failures, disruptions, breaks and other adverse events occurred in the VSC during some period. Consider a group of such random events and denote them by  $E = \{e_1, \dots, e_n\}$ . Basing on the entropy evaluation, we wish to estimate the information (knowledge) about the event occurrences, and, as a result, we wish to forecast which risk types in the supply chains that we can expect in the near future are most hazardous.

For a group of events  $E = \{e_1, \dots, e_n\}$  we assume a priori probabilities of event occurrence  $P = \{p_1, \dots, p_n\}$ ,  $p_i \geq 0$ , such that  $p_1 + \dots + p_n = 1$ , be known. Then the entropy function  $H$  is defined as follows:

$$H = - \sum_{i=1, \dots, n} p_i \log p_i. \quad (1)$$

Let  $R_v$  denote the total number of the adverse events occurred in node  $v$  of the multi-tiered hierarchical SC (see Fig. 2), during the considered time period. If tier  $s$  contains  $n(s)$  nodes, the total number  $E_s$  of adverse events in the tier  $s$  will be  $E_s = \sum_{v=1, \dots, n(s)} R_v$ .

Each adverse event is caused by some factor, which is called a *risk driver*. For the sake of simplicity, we assume that any adverse event is caused by a single driver. Let  $F$  be a total number of drivers; clearly,  $F$  is less than the total number of the registered adverse events in the SC. For each tier  $s$  and each risk driver  $f$  ( $f = 1, \dots, F$ ), using the list of the registered adverse events, we can determine the number  $N_s(f)$  of adverse events caused by

factor  $f$  in all the nodes of tier  $s$ ;  $\sum_{f=1, \dots, F} N_s(f) = E_s$ . Because we are interested in estimating the risk impacts, we introduce another class of random events, denoted by  $A(f, s)$ , which are derived from the initial set of the registered adverse events:

$$A(f, s) = \{\text{the driver } f \text{ is the cause of different adverse events in the nodes of tier } s\}.$$

Denote the probability of  $A(f, s)$  by  $p_s(f)$ .

Using the list of registered adverse events, we can compute the relative frequency  $N_s(f)/E_s$  of the event  $A(f, s)$  and treat this value as an estimation of the probability  $p_s(f)$  of the event  $A(f, s)$ , i.e.,

$$p_s(f) = N_s(f)/E_s. \quad (2)$$

Obviously,  $\sum_f p_s(f) = 1$ .

For simplicity, we assume that the considered events  $A(f, s)$  are independent, and that the loss due to the risks in the SC is summed up from the losses caused by different risk drivers in the system components.

Now we can turn to the computation of the SC entropy. The values  $p_s(f)$ , defined by (2) are taken as a priori probabilities of the events participating in the calculation of the entropy function (1). We presume that the information content may be measured by the entropy at any tier, but the entropy value (and the information) varies from tier to tier because it reflects the contribution of different components (nodes) entering a tier. Recall that each tier describes the same SC, but with a different degree of sharpness.

Consider the entropy  $H(s)$  at tier  $s$  of the SC (the latter will be called below the entropy *for the  $s$ -truncated supply chain*), which is defined as follows:

$$H(s) = - \sum_{f=1, \dots, F} p_s(f) \log p_s(f).$$

Notice that the summation is taken over all the risk types in the considered SC. Now we can observe a non-trivial pattern in the behavior of the entropy of the  $s$ -truncated supply chain.

*Entropy's dynamics.* Consider an arbitrary vaccination SC and let the tier number  $s$  be monotonically increasing:  $s=1, 2, \dots$ . If the set of unknown risk factors shrinks when the  $s$  value grows, the entropy  $H(s)$  of the  $s$ -truncated structure decreases.

This result follows from the following two observations.

First, consider the root-node at tier  $s=1$  (see Fig.2). Although the general number of the risk drivers  $F$  is assumed to be known for the registered adverse events, at this level the probabilities  $p_s(f)$  for all  $f$  are yet unknown to the decision maker. Therefore, in the worst case, they are considered uniformly distributed for all the drivers, and the entropy at this tier is the following:

$$H(1) = - \sum_{f=1, \dots, F} p_s(f) \log p_s(f) = - \sum_{f=1, \dots, F} (1/F) \log (1/F) = \log F,$$

which is the maximum possible value and, hence, reflects the minimum of our knowledge about the risk type distribution in the SC.

Second, as far as we compute the entropy  $H(s)$  for each tier  $s$  moving top down from one tier to the next one, our knowledge about the probabilities  $p_s(f)$ , for different  $f$ , increases, the probabilities cease to be uniformly distributed as they become more diverse, and, hence, more far from  $1/F$ . As a result, the entropy decreases with the growth of  $s$ . This claim was confirmed through extensive experiments.

Taking the observed entropy's behavior into account, we can define which degree of resolution is sufficient when we sequentially compute the entropy for the SC, tier after tier, starting with the root-node. Namely, this computation process can be terminated when the entropy decrease gradient becomes sufficiently small while moving to the next tier. Formally, we appoint an *allowed inaccuracy level*, denoted by  $\varepsilon$ , and may stop the computations at that tier  $s_0$  for which

$$(H(s_0) - H(s_0+1))/H(1) < \varepsilon. \quad (3)$$

For example, in real-life computations in the HCO CLALIT, for the SC fragment presented in Fig.3, we obtained the following results.



The number of the tier	s=1	s=2	s=3	s=4
The number of entities	1	8	26	1,262
The entropy value	H(1)=4.28	H(2)=4.01	H(3)=3.61	H(4)=3.56

We observe that  $(H(4) - H(3))/H(1) = 0,023$ . Therefore, if the allowed inaccuracy level  $\epsilon = 3\%$  then we may stop the entropy calculations at level  $s = 3$ . In practical terms, it means that it is sufficient to study the given SC model limiting ourselves to three tiers only, whereas the study of a larger model with four tiers, that is, 1, 262 nodes can add only 3% of additional knowledge about the risks.

In our experiments in CLALIT with real-life data, we steadily observed the same effect, showing that almost all essential information (knowledge) about risks in the SC can be found within several upper tiers of the model.

## 5. Mathematical model of the cost-benefit optimization problem

Following the logic of our modeling process described in Fig.3, in this section we perform the cost-benefit analysis of the SC on the reduced network model containing reduced number of tiers. The objective is to minimize the total costs. The objective function includes the following four components:

- Delivery costs from the manufacturers and inventory holding costs in the DC.
- Delivery and inventory holding costs within clinics  
Both 'a' and 'b' are direct non-medical costs.
- The shortage cost dependent on customers' waiting time. This cost includes additional medical costs associated with the service of infected people who were not-vaccinated at earlier stages of the vaccination season.
- The service cost proportional to the amount of time required for nurses and physicians to administer the vaccine, multiplied by the average salary per time unit.

The constraints include the material balance in the DC and clinics, the predicted vaccine demand in each time period, the capacity limitations, and the requirements to the medical service level for different population groups.

From the analytical point of view, the model suggested below is a mixed integer programming problem. It differs from the standard inventory/delivery management problem in several ways. First, we consider three echelons of vaccine distribution and utilization: the DC, clinics and population groups. Second, we introduce constraints that impose the required level of service and customer satisfaction. And, third, we explicitly introduce the time index in the supply network model reflecting time-varying costs, benefits and demands.

*Indices:*

$\bar{j}$ - Customer sub-group index,  $\bar{j} = 1, \dots, G$

$i$ - Clinic index,  $i = 1, \dots, I$

$t$ - Time index,  $t = 1, \dots, T$

*Parameters:*

$d_{ijt}$  – demand for vaccine units (doses) in period  $t$  by sub-group  $j$  in clinic  $i$

$h_i^{CL}$  – clinics' inventory holding cost per vaccine unit per time period in clinic  $i$

$h^{DC}$  –DC's inventory holding cost per vaccine unit per time period

$\Pi_{jt}$  – HCO shortage cost per unvaccinated customer in sub-group  $j$  in period  $t$

$A_{ij}$  – service cost per customer in sub-group  $j$  by clinic  $i$

$k^{DC}$  – cost of purchasing and transporting a vaccine unit from the manufacturer to the DC

$k_i^{CL}$  – cost of transporting a vaccine unit from the DC to clinic  $i$

$\alpha_{jt}$  - minimum service level for sub-group  $j$  at time  $t$

$C_t^{DC}, C_{it}^{CL}$  – storage capacity at DC and in clinics, respectively, in period  $t$

$n_0$ - average medical personnel treatment time per customer

$N_{it}$ - medical personnel hours available in clinic  $i$  at period  $t$  (hours/day)

$TQ$ - total quantity of vaccines purchased from the manufacturer.

Decision variables:

$I_t$  – Inventory in DC at the end of period  $t$ ,  $t = 1, \dots, T$

$I_{it}$  – Inventory in clinic  $i$  at the end of period  $t$ ,  $i=1,\dots,I$ ,  $t = 1, \dots, T$

$q_t$  – Delivery (shipping) quantity of vaccine units from manufacturer to DC in period  $t$

$q_{it}$  - Delivery quantity of vaccine units from DC to clinic  $i$  in period  $t$

$s_{ijt}$  –Shortages in sub-group  $j$  in period  $t$  in clinic  $i$

$w_{ijt}$  –Consumption of vaccines linked to group  $j$  in clinic  $i$  in period  $t$

**The min-cost problem:** Minimize the total costs

:

$$J_{HCO} = \sum_{t=1}^T k^{DC} q_t + \sum_{t=1}^T h^{DC} I_t + \sum_{t=1}^T \sum_{i=1}^I k_i^{CL} q_{it} + \sum_{t=1}^T \sum_{i=1}^I h_i^{CL} I_{it} + \sum_{t=1}^T \sum_{i=1}^I \sum_{j=1}^G \Pi_{jt} s_{ijt} + \sum_{t=1}^T \sum_{i=1}^I \sum_{j=1}^G A_{ij} w_{ijt}$$

subject to:

DC inventory balance:

$$(1) \quad I_{t-1} + q_t = I_t + \sum_{i=1}^I q_{it}, \quad t = 1, \dots, T$$

Clinic's inventory balance:

$$(2) \quad I_{i,t-1} + q_{it} + \sum_{j=1}^G s_{ijt} = I_{it} + \sum_{j=1}^G d_{ijt}, \quad i = 1, \dots, I, \quad t = 1, \dots, T$$

Consumption balance:

$$(3) \quad w_{ijt} = d_{ijt} - s_{ijt}, \quad i = 1, \dots, I, \quad t = 1, \dots, T, \quad j = 1, \dots, G.$$

The minimal service level to sub-group  $j$  by time  $t$  in terms of vaccine availability (the minimal ratio of aggregated consumption divided by aggregated demand over all clinics and all periods):

$$(4) \quad \frac{\sum_{i=1}^I w_{ijt}}{\sum_{i=1}^I d_{ijt}} \geq \alpha_{jt}, \quad j=1,\dots,G, \quad t=1,2,\dots,T$$

The quantity of vaccines delivered from the manufacturer to the DC is restricted to the total quantity purchased:

$$(5) \quad \sum_{t=1}^T q_t = TQ$$

Storage capacity constraints in the DC and in the clinics, respectively:

$$(6) \quad I_t \leq C_t^{DC}, \quad t = 1, \dots, T$$

$$(7) \quad I_{it} \leq C_{it}^{CL}, \quad i = 1, \dots, I, \quad t = 1, \dots, T$$

Medical personnel resource availability in clinic  $i$ :

$$(8) \quad \sum_{j=1}^G n_0 w_{ijt} \leq N_{it}, \quad i = 1, \dots, I, \quad t = 1, \dots, T$$

All variables are non-negative.

We solved the above mathematical programming problem for real-life data using the commercial software GAMS ([www.gams.com](http://www.gams.com)).

## 6. Case study

We have applied the mathematical model from Section 5 retroactively to the data for the CLALIT Health Services, the largest HCO in Israel. We obtained a dramatic decrease in the vaccination supply chain size. Whereas the initial four-tier supply chain constructed for real-life data for year 2013 contained about 1,200 clinics located in different geographical places serving about one million people grouped into 6,000 population clusters, the reduced (3-tier) supply chain contained only 26 clinic clusters and 120 population sub-groups.

The optimization model of Section 5 has been formulated on the reduced SC with 3 tiers and applied retroactively for the 2013 data. The mathematical programming problem has been solved by the GAMS software and permitted to decrease the annual total costs by 12%.

## 7. Concluding remarks

This paper develops an integrated approach for optimizing costs and public benefits of an influenza-vaccine supply chain. The approach includes three stages which distinguish it from earlier known works. At the first stage we identify the most significant risk factors and most vulnerable components in a hierarchical vaccination SC using the entropy approach. As a result, at the second stage, we select most meaningful tiers and reduce the entire supply chain model. At the third stage, we formulate and solve the optimization problem aimed at reducing the total costs in the reduced VSC. We believe that the suggested methodology is applicable for other types of healthcare supply chains.

In our future research we intend to accomplish a more scrupulous analysis of links between entropy and costs in medical supply chains. Besides, we intend to perform a more sophisticated cost-benefit-risk analysis of the vaccination SC taking into account stochastic behavior of demands for different population groups.

## Acknowledgement

The research of the fourth author is supported by the grant of Russian Humanitarian Scientific Fund (no. 14-12-40003a(p)). We thank the anonymous reviewers for their constructive suggestions.

## References

- Allesina S., Azzi A., Battini D. and Regattieri A. (2010) "Performance measurement in supply chains: new network analysis and entropic indexes." *International Journal of Production Research*, 48(8), 2297–2321.
- Chick, S.E., Mamani, H. and Simchi-Levi, D. (2008) "Supply chain coordination and influenza vaccination", *Operations Research*, 56, 1493-1506.
- Colombo, G.L., Ferro, A., Vinci, M. and G. Serra, G. (2006) "Cost-benefit analysis of influenza vaccination in a public healthcare unit", *Therapeutics and Clinical Risk Management*, 2, 219-226.
- Durowoju O.A., Chan H.K. and Wang X. (2012) "Entropy assessment of supply chain disruption." *Journal of Manufacturing Technology Management*, 23(8), 998 – 1014.
- Gerdil, C. (2003) "The annual production cycle for influenza vaccine", *Vaccine*, 21, 1776-1779.
- Hebron A., Levner E., Hovav S. and S. Lin (2012) "Selection of most informative components in risk mitigation analysis of supply networks: An information-gain approach." *International Journal of Innovation, Management and Technology*, 3(3), 267–271.
- Isik F. (2010) "An entropy-based approach for measuring complexity in supply chains", *International Journal of Production Research*, 48(12), 3681–3696.
- Kogan K. and Tapiero C.S. (2007) *Supply chain games. Operations management and risk valuation*. Springer. Berlin.
- Levner E. and Proth J.-M. (2005) "Strategic management of ecosystems: A supply chain perspective." In *Strategic management of ecosystem*, edited by E. Levner, I. Linkov, J.-M. Proth. Springer, Berlin, pp. 95–107.
- Matinrada N., Roghaniana E., Rasib, Z. (2013) "Supply chain network optimization: A review of classification, models, solution techniques and future research", *Uncertain Supply Chain Management*, 1, 1–24.
- Nichol, K.L. (2008) "Efficacy and effectiveness of influenza vaccination", *Vaccine*, Vol. 26, pp. 17–22.
- Olson D.L. (2012) *Supply chain risk management: Tools for analysis*. Business Expert Press, New York.
- Shannon, C.E. (1951). "Prediction and Entropy of Printed English". *Bell System Technical Journal* 30 (1): 50–64. .
- Stone, J.V. (2014) *Information Theory: A Tutorial Introduction*, University of Sheffield, England.
- Tang O. and Musa S. N. (2010) "Identifying risk issues and research advancements in supply chain risk management", *International Journal of Production Economics*, 133(1), 25–34.

## **Appendix. List of acronyms**

CLALIT –the name of the healthcare service organization in Israel

GAMS – the name of a commercial software package

DC – distribution center

HCO – healthcare organization

SC -supply chain

VSC – vaccination supply chain

# Google Analytics as a Prosumption Tool for Web Analytics

Joanna Palonka<sup>1</sup>, Marcin Lora<sup>2</sup>, Tomasz Oziebło<sup>1</sup>

<sup>1</sup>Department of Informatics, Faculty of Informatics and Communication, University of Economics in Katowice, Katowice, Poland

<sup>2</sup>Asseco Poland S.A., Katowice, Poland

[joanna.palonka@ue.katowice.pl](mailto:joanna.palonka@ue.katowice.pl)

[marcin.lora@gmail.com](mailto:marcin.lora@gmail.com)

[t.ozieblo@gmail.com](mailto:t.ozieblo@gmail.com)

**Abstract:** The unique character of modern ICT has been formed by the ongoing process integrating sound, images, graphics and numerical data on one media platform – the Internet. Due to its versatility, ICT solutions based on this medium are commonly applied by companies in multiple business operations. The paper focuses on marketing activities. Each organisation which wants to reach new customers, retain the existing ones and gain a competitive edge over competitors must have some tools at hand in order to study particular behaviour of its website visitors. An analysis of Internet data allows for comprehensive monitoring of website operations, in particular the goals defined for the website as well as the needs of the organisation, its customers and business partners. It enables companies to create and optimise online marketing strategies, including e.g. promotion, sales, etc. In Poland, web analytics and the tools which support it are still hardly known and used. The aim of the paper is to verify a hypothesis that Google Analytics (GA) is a prosumption tool for Internet data that can be used by a company to effectively (in a simple way and on its own) manage its website, i.e. manage the website content and traffic. To test the hypothesis, an experiment was designed and conducted on the website of Soluzioni IT from 1 May 2012 to 31 May 2013. The experiment was divided into four stages. After implementing GA on Soluzioni.pl site, the basic parameters were configured, the major goal (to increase the website traffic) and the particular ones were determined and Key Performance Indicators (KPIs) were assigned to them. During the first stage, a preliminary analysis of the site was made. The values of KPIs obtained from GA at this stage were adopted as a reference point for the KPIs of the subsequent stages. The scientific experiment performed on the website involved active modification of its content and graphic design. The GA indicators were monitored, the impact of the changes on the GA values was assessed and subsequent changes were designed and implemented. The process was cyclically repeated in the next stages of the study. It allowed for the identification of the causal relationship between the website content and the visitors' behaviour. The observation of the website traffic parameters, which was part of the experiment, contributed to determining the areas of high potential for the website traffic optimisation. The results of the experiment confirm the hypothesis that GA is an effective tool for optimising the management of the corporate website content and traffic. The users themselves can compare the data they need, analyse goal conversion and take decisions concerning the online activities. Web Analytics and the Internet data analysis tools can be regarded as modern tools supporting the management of an organisation.

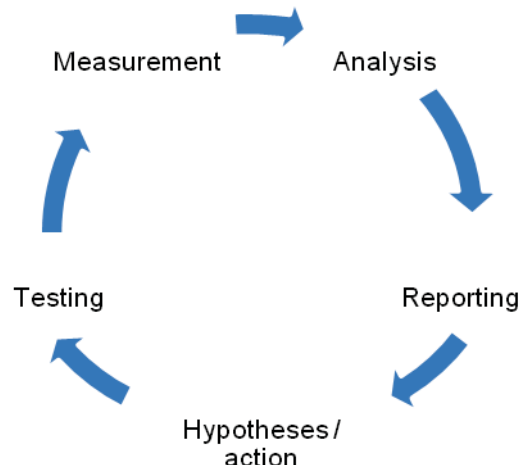
**Keywords:** e-marketing, Web Analytics, prosumption tool, Google Analytics, website traffic optimisation, website content management

---

## 1. Introduction

Web Analytics involves traffic monitoring and visitor behaviour tracking on a particular website with the aim of optimising an organisation's marketing strategies (Kaushik, 2009). Web Analytics activities are conducted cyclically (Figure 1) and must not be implemented piecemeal. They can be divided into 5 stages, which have to be carried out in accordance with the following principles (Kowalczyk, 2012):

1. Combine analytics with business.
2. Share knowledge about analytics.
3. Customise reports for different recipients in an organisation (an analyst, a supervisor, HIPPO)
4. Set business objectives and indicators.
5. Measure and analyse conversion.
6. Segment data.
7. Integrate off-line and on-line data.



**Figure 1:** Web analytics as a continuous process

Measurement is the first stage of Internet data analysis. It involves defining website objectives (strategic and marketing) which an organisation is going to achieve (Lewiński,2013). Strategic objectives should be doable, understandable, manageable and beneficial. Next, KPIs for the entire project should be defined (Peterson,2004),(Lovett,2011) and translated into the language of statistics tools. Thus a decision of what is going to be measured during the analysis of a particular site is made. At this stage, an organisation has to determine what values of particular indicators will be considered satisfactory in a specific period of time (Clifton,2008) (Table 1).

**Table 1:** Examples of KPI plans

Strategic objectives	Marketing objectives	KPI	KPI targets
Product sales	Profit increase	Total Revenue	PLN 20,000/month
	Sales volume increase	Visit value	PLN100 Per visit Value
Effective internet marketing	Creation of a prospect database	Number of new visitor registrations	1,000 registrations/month
Brand building	Rise in the willingness to buy	Conversion Rate	2%
	Increased involvement	Time on Site	> 1 minute

During the second stage – analysis – an analytics tool is chosen and configured. According to the principle *rubbish in rubbish out*, if a mistake is made at that point, even the best analysis will lead to erroneous conclusions (Trzósło,2014). When an account is set up, it must be configured, i.e. it is necessary to determine an account structure and create profiles assigned to users of a selected part of the website, configure goals, create alerts, describe filters, etc. Tools can be combined to obtain more complex data on PPC (pay-per-click) campaigns and free search results, e.g. Google Analytics with Google Adwords (Bailyn,Bailyn,2012),(Clifton,2008).

During the third stage – reporting – website statistical reports are generated. Modern analytics tools offer a wide range of standard reports and enable creation of custom reports as well. There are various possibilities of data visualization (Few,2009). Reports can give answers to different questions, including (Cutroni,2010):

- Where are our users from?
- How much time do they spend on particular pages?
- What percentage of them abandon their shopping cart at the last stage?
- How many Unique Users convert to the site goals?

- Which channel (for example organic positioning, Adwords or a banner campaign) is best, and which has to be immediately given up?
- What do people search for in our internal search engine?
- etc.

However, reports themselves do not suffice. An overall look at the data must be taken to draw conclusions that will be profitable for a company (Burby,2007). The results of analyses can help improve the quality of services, save resources or make some business processes more efficient. Based on the data, business can be improved (Krug,2013). But it is also necessary to assess the path between the data and improvement activities, e.g. those increasing company efficiency. There is one principle – analytics should lead to continuous optimisation. When a report is ready, the data should be analysed, recommendations for changes made and then implemented (this is a key moment of the entire process and the most important stage as it brings a company real profits) (Ash,Ginty,Page,2012).

In stage 5, the progress in goal accomplishment is checked. Usability tests are done, site content and design are tested (on-line tests – A/B or multivariate) and surveys are conducted (Beasley,2013).

Web analytics is carried out by means of IT tools. So far, the analysis and evaluation of Internet data have been conducted by specialised consulting firms. Nowadays, organisations themselves are getting more involved in such operations as a result of employees' increasing knowledge and competence and wider availability of tools that are used for this kind of analyses (Dykes,2011),(Rosenfeld,2011). There are free tools, e.g. Google Analytics, Piwik, Open Web Analytics and paid ones, e.g. Gemius, Coremetrics, Adobe Online Marketing Suite, Webtrends, etc. (Oberoi,2014). IT tools provide valuable data, but making and implementing recommendations are critical to web analytics. The synergy of these elements allows for continuous creation of innovative ideas and higher profits.

Google Analytics (GA) is gaining recognition among Polish entrepreneurs as a tool for the analysis of the global Internet traffic. It is an on-line analytics service which enables users to manage their website, i.e. its content and traffic, on their own. It can be used to measure and assess various aspects of Internet projects and perform comprehensive website analysis. It provides detailed information on web traffic by means of segmentation, custom reports, charts and comparison tools. Its data analysis features offer instant access to a great number of data, metrics, KPIs, etc. (Kaushik,2010).

GA can be used to (Clifton,2008),(Gąsiewski,2013):

- 1) calculate Return on Investment – sophisticated data analysis and reporting features enable companies to find out which keywords, used for example in sponsored links, are the most profitable. The information allows for more effective marketing investment. It is also possible to calculate customer acquisition cost.
- 2) check traffic sources, i.e. determine where exactly users are coming from, not only the geographic location but the referring sites as well. As a result, a company gets information where it should advertise and where its advertisements really pay off.
- 3) check content effectiveness, i.e. identify pages which have a high bounce rate and those whose content is particularly interesting to users. Having such knowledge, an organisation can change the content of pages with a high bounce rate or expose those which appeal to Internet users.
- 4) set goals and create goal paths - for each site which has an implemented GA code, specific goals, e.g. product sales, form completion, reaching a certain page, and a path to completing the goals can be set (Tonkin,Whitemore,Jutroni,2010). Thus an organisation can find out at which stage it loses users and take steps to keep them on the site.

In addition, GA can send customised emails and export data in various formats. Data can be visualised by means of interactive maps and site overlays or presented graphically, as charts (bar charts and pie charts) and tables. The service allows for continuous monitoring and analysis of the website traffic, identification of most frequently viewed pages, etc.

Depending on the site goals, GA answers numerous questions (Figure 2).



**Figure 2:** GA–basic questions

GA answers two questions which are of critical importance to running a business (Kowalczyk,2013):

1. What happens on the website in terms of users' activities?

- Why did you enter my site?
- Did you get what you were looking for?
- If not, tell me why?

2. How to edit a website and what marketing operations to undertake in order to reach goals?

The aim of the paper is to verify a hypothesis that GA is a presumption tool for Internet data analysis that a company can use to effectively (in a simple way and on its own) manage its website, i.e. manage the website content and traffic. Depending on their needs, users can gather any data and next analyse and evaluate them. The results can help streamline the marketing operations in terms of promotion, advertising and sales operations and plan new website goals.

An experiment was performed to test the hypothesis. It was carried out on the website of Soluzioni IT, which provides custom-built IT products and services.

Chapter 2 presents the assumptions and the course of the experiment, while chapter 3 contains the resulting conclusions.

## **2. Google Analytics as a presumption tool for Internet data analysis**

Over the last few years one can observe increased involvement of direct IT system users in activities which used to be outsourced. This happens in business, with regard to IT technologies and tools, which are designed in such a way that users are able to collate data and conduct analyses for their own needs. The IT tools of this kind are called presumption tools (Gajewski,2009),(Szymusiak,Toffler,2013),(Toffler,1997).

The paper attempts to test a hypothesis that GA is a presumption tool which gives a chance to continuously create innovative ideas and realise higher profits. The service makes innovation more likely to succeed since users, not employees of R&D departments and consultation firms, define specific needs. The assumption was that thanks to GA an organisation can effectively (in a simple way and on its own) manage its website, i.e. manage the website content and traffic. The study was carried out on the website of Soluzioni IT. Operations which were undertaken based on the GA analyses were to result in website traffic optimisation, i.e. a rise in



traffic due to improved positioning in search engines and a greater number of users interested in its content. The traffic optimisation effect was to be reflected in an increasing number of visitors, a larger number of new users or a larger number of returning visitors. At the same time, site visitor retention was evaluated and content that would meet visitors' requirements was provided. This, in turn, was to increase the number of pageviews per visit, reduce a bounce rate and increase time spent on the website. As a result, the goals set by the owner were to be achieved.

## **2.1 The course of the study**

The optimisation work on Soluzioni.pl website, by means of GA service, was conducted from 1 May 2012 to 31 May 2013. The website was cyclically assessed and redesigned. The operations were possible owing to proper configuration of measurements, appropriate segmentation and data filtering. Mistakes and shortcomings on the website were cyclically eliminated, which led to a steady increase in traffic. In addition, monitoring of major indicators enabled researchers to understand visitors' behaviour and identify areas of great potential for improved traffic.

The study was carried out in stages comprising:

- preliminary analysis and website assessment – stage I – 1 May 2012– 30 September 2012
- modification of the website content and a change of its graphic design – stage II – 1 October 2012 – 30 November 2012
- website popularisation on other sites – stage III – 1 December 2012 – 28 February 2013
- work on website positioning in the Google search engine - stage IV – 1 March 2013 – 31 May 2013

During the *first stage*, after installing and activating GA, a preliminary analysis of the website was made and the initial values of metrics were determined. In order to accurately identify the project objectives, a general category of the owner's desired objective was determined (Lovett, 2011). The main goal was set – to increase website traffic - and then it was translated into particular goals, i.e. optimisation of the website content and structure according to users' requirements, web positioning and on-line promotion through social media. Next, measurable KPIs were assigned to the goals, including the total number of visits, the number of new visitors, the percentage of new visitors, visits from Google search engine, direct traffic, the average time spent on the website, the average number of pages viewed during a visit, an exit rate. Constant monitoring and control of the KPIs' values ensured proper assessment of the website optimisation effects with respect to the changes taking place over time.

First, the indicators related to the main goal were monitored and later on, those connected with the website content and the website visitors' behaviour.

Consequently, within the Soluzioni.pl profile the following website goals were set and progress towards them was being monitored over time.

*Goal 1* – Visitors moving from the Offer page to the Contact page (the goal for which the custom alert was defined).

*Goal 2* – Visitors spending more than 5 minutes on the site.

*Goal 3* – Visitors not entering the site via the Google search engine.

*Goal 4* – Visitors from the Google search engine – the goal which permits examination of web positioning effectiveness.

Moreover, a path for goal 1 was set up. It was to show a sequence of users' steps on the way to the goal. By monitoring the path it was possible to obtain detailed information about users abandoning the process at particular stages. The information is essential for website evaluation and optimisation.

The completion of *stage I* – a preliminary analysis and evaluation of goal achievement levels – showed that:

- There was little interest in the website as the traffic volume was not high – 133 visits (including 122 UU) and only 281 pageviews
- The site content did not match the visitors' expectations and they left it immediately, which was reflected in the bounce rate = 68.42%, the percentage of returning visitors = 12.03% and the average visit duration of less than 60 seconds. Similarly, the number of pageviews per visit was low – only 15% of the users visited more than 3 pages during one visit.
- The site met technical requirements – the results did not indicate any problems with the site being displayed in various search engines and operating systems.
- Most visitors came to the site from referring sites (42.86%); 32.33% visitors from direct sources (Direct Traffic) and 24.81% used search engines to find the information.
- As for the traffic sources, the most visited site was Homepage (103 views), followed by the Price list page (26 views), the Offer page (21 views) and the Projects page (20 views). The Contact page, which was essential for the accomplishment of Goal 1, was not one of the top ten.
- Among landing pages, Homepage had the greatest number of visits, the second position was held by the Action page, Database and ETL. A bounce rate of the landing page was 3.08% lower than the average, which showed users' willingness to further browse the site. The Action page had a bounce rate of 100%, which might indicate that it lacked the information the visitors were looking for, e.g. via a search engine.
- Most pageviews came from a referral source and the users who came through that source visited Homepage and Projects. Users from the Google search engine also visited Homepage and Trivia, but their number was more than 50% lower.

Summing up the findings of this stage, a conclusion was reached that the biggest problem was site visibility in the search engine (goals 3 and 4 had not been achieved). It was necessary to start working on search engine optimisation (SEO), which would lead to a reduced bounce rate and a drop in the number of visitors leaving the site. Referral sources were pointed out as the right direction for site traffic optimisation and increase. It was decided that the site should be promoted on selected platforms (e.g. facebook.com, peb.pl). Therefore, work had to be undertaken to enhance the chance of the site being found on Google and, at the same time, to increase the number of visitors referred by other sources.

Text modification was another important step towards improving site traffic and search capability. Changes in the content of the site and the pages as well as in the general layout were proposed with the aim of contributing to goal 1 achievement.

In the time period under study, goal 1 was not achieved, which mainly resulted from low website traffic and difficult contact through the Contact page. It was decided that the Contact page had to be altered and displayed in a place visible to users.

The preliminary analysis led to the identification of the areas and pages which needed change. Particular operations were recommended for the subsequent stages of the study to reach the set goals.

During *stage II* changes recommended as a result of the preliminary analysis were implemented. In that period the website content and the overall graphic design were modified.

After completing the work, the level of goal achievement was measured. The results were as follows:

Goal 1 – 21 visitors

Goal 2 – 31 visitors

Goal 3 – 247 visitors

Goal 4 – 56 visitors

The comparison of the data from the first period and the second one shows that progress was made. A thorough analysis indicated that site traffic increased and so did the number of pageviews per visit, the bounce rate fell, users spent more time on the site and the number of potential clients rose due to the fact that goal 1 was attained. In addition, changes in visit sources were observed. It was stated that further work on the site should produce even better results.

**Stage III – website popularisation on other platforms – focused on visit sources and on the popularisation of the site on social networking sites. The study was based on Facebook.com and internet forums. On Facebook social networking site a profile for Soluzioni IT company was created (Figure 3).**



**Figure 3:** 1 Facebook profile of Soluzioni.pl

The aim was to establish communication with potential customers – community members. Moreover, activities were undertaken on various internet forums, which was to attract more visitors from other websites (i.e. referral sources). The promotion of the site in social media was to increase its traffic.

In that period, there were 17 visitors moving from the Offer page to the Contact page (goal 1), 24 visitors spending more than 5 minutes on site (goal 2), 263 visitors not entering the site via the Google search engine (goal 3) and 103 visitors from the Google search engine (goal 4).

Comparing the data from periods I and III, a positive impact of the changes was observed. They brought about more traffic, bigger number of pageviews per visit and longer visit duration.

The number of the goals achieved was similar to the previous period, which indicates the right trend in site development.

The main goal was to increase the number of visitors from referral sources, yet it grew by less than 4%. Therefore, a decision was made to continue the promotion of the services on Facebook.

During *stage IV* attention was focused on positioning the website in the Google search engine. At the same time, the Facebook campaign was continued. The main aim was to introduce changes that would ensure better search engine visibility of the site and reduce the bounce rate.

The measurement of the goal achievement level showed that :

Goal 1 – 17 visitors moved from the Offer page to the Contact page

Goal 2 – 31 visitors stayed on the site for more than 5 minutes

Goal 3 – 285 visitors did not access the site via the Google search engine

Goal 4 – 156 visitors came to the site from the Google search engine.

The comparison of data from period I and period IV showed a dramatic increase in site traffic. The number of visits was the highest ever. In addition, compared to the previous period, more than 50 users entered the site through the Google search engine, which shows how important site positioning is for traffic optimisation.

## 2.2 KPI evaluation

Table 2 presents values of the KPIs for the periods under study and their changes during the study, which gives a reliable assessment of their impact on traffic optimisation and goal achievement.

**Table 2:** The values of KPIs for the time periods under study

	I period	II period	%change	III period	%change	IV period	%change
Total number of visits	133	309	132%	374	21%	455	19%
Number of new visitors	122	235	93%	322	37%	372	16%
Percentage of new visitors	92%	76%	-16%	86%	10%	82%	-3%
Visits from the Google search engine	33	56	67%	103	87%	156	51%
Direct traffic	43	76	77%	138	82%	194	41%
Avg time on site	00:31	02:26	363%	01:32	-37%	01:18	-15%
Avg pageviews per visit	2,11	3,61	71%	3,02	-16%	2,61	-13%
Exit rate	68,42%	44,66%	35%	51,87%	-16%	51,69%	0,36%

The values for *the total number of visits* show that the changes inspired by the GA findings led to a significant increase in the site traffic. The number of visits kept rising steadily to become almost four times bigger than in the first period of time. The most dramatic growth (of 132%) could be observed after the first site optimisation (period II). This was accompanied by a considerable increase in the number of new visitors (from 122 in the first period to 372 in the last one).

*The percentage of new visitors* was falling (except for period III). However, the drop was of a few percent only and thus of minor significance - the biggest change could be observed between the first two time periods (-16%) due to building a relatively strong base of returning users.

*The average Time on Site* in the first time period rose by 363%, which might result from the new layout of the site. However, after that, the visits were getting shorter to finally level off at 1:15. The figures could mean that the site content was well structured and the users visiting it were able to quickly find the information they needed.

*Visits from the Google search engine* and *direct traffic* rose in each period. This proves that the adopted optimisation strategy, also based on the GA data, was effective. An increase in the number of visits from the Google search engine shows the effectiveness of positioning and makes it an essential method for boosting online visibility.

Throughout the whole period of time, *the average number of pageviews per visit* ranged from 2 to 3 pages and cannot be regarded as a satisfying score. Due to the modification of the site content, *the exit rate* almost halved compared to the first period, which is quite a good result.

In summary, the analysis of the GA data helped to achieve all the goals set at the beginning of the study. Gradually, the problems on the website were eliminated, new marketing operations were implemented and the site content was optimised. As a result, site traffic increased and the scope of site access was extended. As the data in table 2 show, owing to the site modifications the values of all KPIs improved.

In most cases, the values did not rise sharply between the time periods. However, in a few cases, an increase of more than 100% compared to the previous period could be observed, e.g. a growth in the total number of visits between the first two periods, which shows how effective the undertaken site optimisation was.

### 3. Conclusions

Web Analytics can be used to manage the implementation of an organisation's strategy. By undertaking specific activities, it focuses on the achievement of particular goals. The decisions regarding those activities are based on the knowledge acquired from various analyses. Moreover, it provides the ability to monitor the extent to which the goals are being achieved. Such a view indicates that web analytics should be well-organized and structured, and its processes should be of continuous character. This can be attained with the

help of IT. The result of the experiment confirmed the hypothesis that GA is a presumption tool for Internet data analysis. It allows owners to comprehensively monitor the parameters of any website. The software can track users' path on Internet platforms and visualise the results, e.g. generate a map of mouse movements and clicks. Thanks to GA results, it is possible to precisely target advertisements, use various means to strengthen marketing initiatives and design websites which can generate more conversions. GA provides information on how users find a site and what their interactions are. It enables comparison of changes in users' behaviour due to e.g. improvements in site content and design. It is a good analytics tool which ensures comprehensive analysis of the whole site from multiple perspectives, depending on users' needs. However, the data themselves don't suffice to effectively support the management of the online operations. Using GA requires the ability to interpret data from individual reports and identify relationship between particular data as well as knowledge about products, services and customers' needs. A lack of such skills and knowledge can lead to wrong decisions which will negatively affect website goals.

Summing up, Web Analytics can be regarded as a modern instrument in the management of the online operations. Google Analytics is a supporting tool which supplies detailed information about, i.a. the behaviour of visitors to a particular website, the usability of the content provided, etc. The synergy of those elements is able to effectively support the decision-making process and optimise management activities. The above conclusions can provide the basis for creating a new way of managing the organisations that conduct online operations, which implies the necessity for further research.

## References

- Ash, T. and Ginty, M. and Rage, R. (2012) *Landing Page Optimization: The Definitive Guide to Testing and Tuning for Conversions*, Sybex.
- Bailyn, E. and Bailyn, B. (2012), *Przechytrzyć Google. Odkryj skuteczną strategię SEO i zdobądź szczyty wyszukiwarek*, Helion, Gliwice.
- Beasley, M. (2013) *Practical Web Analytics for User Experience*, [ebook], Morgan Kaufmann Publishers.
- Burby, J. (2007) *Actionable Web Analytics: Using Data to Make Smart Business Decisions*, Sybex.
- Clifton, B. (2008), *Advanced Web Metrics with Google Analytics*, John Wiley&Sons.
- Cutroni, J. (2010) *Google Analytics*, O'Reilly Media.
- Davenport, T.H. and Morison, R. and Harris, J.G. (2010) *Analytics at Work: Smarter Decisions, Better Results*, Harvard Business Review Press.
- Dykes, B. (2011) *Web Analytics Action Hero: Using Analytics to Gain Insight and Optimize Your Business*, AdobePress.
- Few, S. (2009) *Now You See It: Simple Visualization for Techniques for Quantitative Analysis*, Analytics Press.
- Gajewski, Ł. (2009) „Prosumpcja – praktyki konsumenckiej innowacyjności”, [online], <http://www.e-mentor.edu.pl/artykul/index/numer/29/id/631>.
- Gąsiewski, M. (2013) „Przewodnik po Google Analytics”, [online], [www.ittechnology.us/ebook-google-analytics/](http://www.ittechnology.us/ebook-google-analytics/).
- Kaushik, A. (2009) *Godzina dziennie z Web Analytics. Stwórz dobrą strategię e-marketingową*, Helion, Gliwice.
- Kaushik, A. (2010) *Web Analytics. Świadome rozwijanie witryn internetowych*, Helion, Gliwice.
- Kowalczyk, D. (2012) „Najważniejsze zasady analityki internetowej”, [online], [www.damiankowalczyk.pl/2012/05/najwazniejsze-zasady-analityki.html#.UZ--TpyjvDg](http://www.damiankowalczyk.pl/2012/05/najwazniejsze-zasady-analityki.html#.UZ--TpyjvDg).
- Krug, S. (2013) *Don't Make Me Think, Revisited; A Common Sense Approach to Web Usability*, New Riders Publishing.
- Ledford, J. and Tyler, M. (2009) *Google Analytics 2.0*, Wiley Publishing, Inc.
- Lewiński, M. (2013) *Google Analytics, Materiały szkoleniowe Akademia Analityki*, Warszawa.
- Lovett, J. (2011) *Social Media Metrics Secrets*, John Wiley&Sons, New York.
- Oberoi, A. (2013) “Top 30 Web Analytics Tools”, [online], [www.adpushup.com/blog/web-analytics-tools-google-analytics-alternatives/](http://www.adpushup.com/blog/web-analytics-tools-google-analytics-alternatives/).
- Peterson, Eric T. (2004) *Web Analytics Demystified: A Marketer's Guide to Understanding How Your Web Site Affects Your Business*, Ingram.
- Rosenfeld, L. (2011) “Search Analytics for Your Site”, [online], <http://rosenfeldmedia.com/books/search-analytics/>.
- Tonkin, S. and Whitmore, C. and Jutroni, J. (2010) *Performance Marketing with Google Analytics: Strategies and Techniques for Maximizing Online ROI*, Wiley Publishing, Inc.
- Toffler, A. (1997) *Trzecia fala*, PIW, Warszawa.
- Szymusiak, T. and Toffler, A. (2013) *Prosument czy Lead User współczesnej ekonomii?*, Creativetime, Kraków.
- Trzósło, T. (2014) „Analityka internetowa”, [online], <http://semkonsultant.pl/uslugi/analityka-internetowa/>.

# **Analytics in Context: Modelling in a Regulatory Environment**

**Albert Simard**

**Integrated Knowledge Services**

**Ottawa, Ontario, Canada**

[albert.simard@outlook.com](mailto:albert.simard@outlook.com)

**Abstract:** This paper describes a framework for managing the modelling process in a regulatory environment. It positions analytics as the core of a broader process that begins with a decision to employ modelling to address a regulatory issue and ends with using model outputs to support regulation in the real world. The paper begins by outlining systems, modelling, and framework concepts that underlie modelling. It then discusses content, in the form of data, information, and knowledge that define and describe a model. A management decision guide, based on these underlying concepts, comprises four phases – demand (new model needed), supply (existing model available), project (model development), and application (using model outputs). Each phase consists of two to four stages which, in turn, include a number of steps and considerations. Each step involves an explanation, answering a question, writing a brief statement, and selecting the next step. The decision guide emphasizes appropriateness for an intended use, facilitates interaction among modellers, managers, and users, and documents due diligence for resisting challenges to the model. The framework eliminates wasted time, expense, and effort of developing a model that does not, ultimately, address an intended issue. It also reduces the chance of misusing an existing model that does not adequately represent a situation.

**Keywords:** Modelling, framework, decision guide, regulation

---

## **1. Introduction**

Regulatory agencies manage complex natural, economic, and social systems for the overall benefit of society. For example, the Canadian Food Inspection Agency (CFIA) intervenes in the food sector to safeguard food, animal, and plant resources and their products. Regulated systems are typically complex which generally precludes directly observing or predicting the outcome of intervention alternatives. By simplifying reality, models provide a method for integrating and analysing the many diverse components of the food safety system. Models can shift the balance from using available but incomplete and uncertain results to using robust, transparent, evidence-based results to support decision making. The range of existing and potential uses for models is substantial. Examples include multi-criteria decision analysis and food safety (Frazil et al, 2008), regulatory models and the environment (Holmes et al, 2009), models in environmental regulatory decision making (National Academy of Sciences, 2007), and risk analysis (Greenberg and Cox (2012).

The food safety system spans a broad and diverse spectrum of processes ranging from “farm to fork” (Duram, 2006). The UN-WHO and FAO (2006) provide international standards for food safety. Food safety includes understanding natural process such as the onset and spread of diseases (biology, meteorology, and epidemiology), to optimizing operating practices (agriculture, transportation, and food processing), through the socioeconomic consequences of regulation (communication, profitability, competition, and trade), and finally to managing unpredictable events (risk management, detection, response, mitigation, and recovery). Appropriate modelling approaches differ substantially across this broad range of needs. It is important, therefore, to understand how the conceptual underpinnings of these needs and their underlying content affect modelling. It is equally important to understand the limitations of models in the real world and to use them within their limits. For example, Garner et. al. (2007) described the disastrous consequences of the inappropriate use of models for managing an outbreak of foot and mouth disease in England.

A government regulation is a rule or law designed to control or govern social or business activity; it generally includes monitoring and legal enforcement of the rules. Regulations create, limit, or constrain a right, create or limit a duty, or allocate a responsibility. While regulations benefit society as a whole, they are contentious because compliance is mandatory and it usually reduces productivity and/or increases costs. Further, regulations are based on values as much as analysis (e.g., the value of life); the question of how much is necessary is often hotly debated. Therefore, any weakness in a regulatory model will be challenged by those who are adversely affected by its outputs. As an abstraction of reality, no model is perfect. Further, modelling is part science, part computer programming, part analysis, part interpretation, and part art. As a result, the design, development, and implementation of a model must be appropriate, transparent, documented, defensible, and acceptable to stakeholders.



Regulatory modelling must also consider more than the technical modelling process. Managers focus on the time, effort, and cost of model development which are, in turn, driven by externalities such as the situation, resource availability, and the budget. In contrast, regulators focus on appropriateness of the model, consequences of error, and the socioeconomic implications of implementation. Modellers, managers, and regulators work in very different cultures and see things quite differently. Consequently, a key purpose of this modelling framework and decision guide is to provide a common platform to support conversations among these three groups.

This paper summarizes a modelling framework and decision guide developed for the CFIA (Simard, 2009, see Figure 4. It begins by considering underlying concepts of modelling (systems approach, model development, and framework attributes). It then describes the content (data, information, and knowledge) that define and describe a model. This is followed by an outline of a modelling decision guide that applies the underlying concepts through four phases of the modelling process (demand, supply, project, and application).

## **2. Underlying Concepts**

Concepts are examined from three perspectives. First, because uncertainty and dynamic behavior are inherent in complex systems, systems analysis principles are described in a modelling context. Second, modelling is described as an iterative feedback process involving interactions among modellers, managers, regulators, and stakeholders. Finally, desirable framework attributes are described.

### **2.1 Systems**

The idea that the whole is greater than the sum of its parts is the essence of what differentiates the systems approach from the traditional scientific method. The systems approach analyzes large numbers of components, relationships, and behavior as a whole. A system is a group of interrelated components, within a structure, under common control, that use resources to transform inputs into outputs, to achieve a common goal.

System structure includes an environment that provides resources and receives outputs, a boundary between the system and its environment, a multi-level hierarchy, and interactions among many components. All open systems – whether human or natural – require resources, in the form of matter and energy to function and sustain themselves. A flow of information to and from all components is essential for a system to function collectively.

Systems can be classified into four levels, based on their overall behavior and complexity.

- *Flow-through systems* have simple, predictable behavior. Models are based on well-defined mechanistic input-output processes. Decisions are based on facts with little interpretation.
- *Feedback systems* have complicated linear behavior. They can incorporate memory, anticipation or delays, and can modify their outputs. Mathematical models are typically used for analysis. Decisions require professional knowledge and interpretation.
- *Learning systems* have complex nonlinear behavior. They can change their goals in response to or anticipation of changing environmental conditions. Simulation models are typically used for analysis. Decisions require substantial experience and judgement.
- *Emergent systems* have unpredictable behavior. They can result in sudden transformations or jump-shifts with new properties that are irreversible. Systems with emergent properties would rely on “what if” scenario analysis. Decisions involve acting, sensing, and adapting.

Modelling involves all four categories of systems. Some stages, such as data management and model evaluation, primarily involve simple linear flow. Many stages, such as model development and model evaluation involve feedback and multiple iterations as they spiral towards a solution. Other stages, such as project management and implementation involve learning and goal changing. Finally, a model may create emergent knowledge that results in totally new forms of regulation. A systems approach was used to develop the decision guide.

## **2.2 Modelling**

A model is an abstract and simplified representation of reality that enables regulators to better integrate and understand natural, economic, and social systems of interest. Modelling is the process of developing, managing, and using models to support regulatory activities. The framework includes an inventory of existing models coupled with a capacity to develop and adapt new models. There are three perspectives from which to view modelling, those of: modellers, managers, and users. Managers provide authority and allocate resources to develop or adapt models; modellers undertake the technical work of model development; while regulators determine the nature and extent of their use.

A set of principles should guide model development in a regulatory environment.

- Development effort should not exceed the importance of the intended use.
- Every model element should increase output quality for the intended use.
- A model should not be extrapolated beyond its knowledge base.
- A model should not exceed the availability of relevant input data.
- Stakeholders should be involved in development and endorse a model.
- A model should be accessible to enable external evaluation and use.
- A model should emphasize understandability over elegance or sophistication.
- A model should not be used beyond the limits of its designed applicability.

In essence, there should be balance in all aspects of the modelling process – between completeness, simplicity, and effort; between knowledge, data, and design; between transparency, accessibility, and intended use. Following the principles increases the applicability and usefulness of models and modelling. However, managing the modelling process is notably less advanced than model development. Although the decision guide addresses this issue, the skill, judgement, and experience of the modeller are key to efficient modelling and relevant outputs. The modelling principles are incorporated throughout the modelling decision guide.

## **2.3 Framework**

Like a blueprint or architectural drawing, a framework shows how a complicated structure is put together, the relationships among its parts, and it helps to understand how it works. It also provides a foundation for planning and decision making. A framework lists everything needed to implement a strategy or program. It also provides an outline for planning programs, projects, and activities. By linking the organizational mandate, legal obligations, policy directives, and standards, a framework promotes consistency at different organizational levels and across different functions and business units.

Attributes of the modelling framework include:

- *Independent* of content, issues, situations or the organizational structure.
- *Applicable* to any food safety assessment, management, or communication issue.
- *Perspectives* of managers, developers, regulators, and stakeholders are considered.
- *Scalable* upwards to multiple partners or downward to individual projects.
- *Primary driver* is emerging issues and demand-driven needs.
- *Secondary driver* is using existing models and supply-driven needs.
- *Process* emphasizes situational needs rather than technical sophistication.

## **3. Content**

This section extends overall conceptual underpinnings of systems, modelling, and frameworks to the underlying content that defines and describes a model. Content is a collective term that includes data, information, and knowledge, all of which are essential to modelling. Although often misused or interchanged,



the three forms of content are qualitatively different and it is important to understand the nature and distinct role of each in the modelling process. Attributes of and uses for the three forms of content are described in this section.

### **3.1 Data**

Data are recorded, ordered symbols or signals that may carry information and patterns that describe or define a concept or object. A model and its input data are inseparable; they succeed or fail as one. It is as important to get the data right as it is to get the model right. Data acquisition describes how data needs, data categories, and scale relate model inputs to the intended use of a model. Model data are considered in the data stage of both supply and demand phases of the decision guide.

Many models need data on natural processes, which tend to be measurable with a high degree of confidence. However, models of natural processes must be interpreted in terms of system impacts or intervention outcomes. System models need data about the food safety system, which are more difficult to measure and may have a higher level of uncertainty than data on natural processes. Finally, intervention models need data that enable estimating potential socioeconomic outcomes of intervention alternatives. These data are not directly measurable and have the highest level of uncertainty. Despite the uncertainty, the ability to model complex system processes and multiple intervention outcomes normally leads to better decisions than depending on human experience and judgement.

Urgent situations need real-time data, which means automatically resolving: errors, missing data, access delays, and system malfunctions. Past events need historical data, which are less challenging to acquire but only previously collected data are available, which may not match current needs. Proposed actions need future data, which pose the fewest challenges because exactly the right data can be collected. Static data describe the state of a situation at a particular time, in contrast to dynamic data which describe trends or changes of state over time.

Data must be collected at spatial, temporal, or process scales that are appropriate to the situation. Too much detail wastes time, resources, and effort while too little detail may conceal processes that are essential to obtaining useful results. One-way transitions between adjacent scales may be possible. While integration is generally feasible, differentiation into component parts is normally impossible.

### **3.2 Information**

Information is meaning in context arising from processing, interpreting, or translating data to identify an underlying message or pattern. Information systems integrate people, work processes, governance, and technology to acquire, store, process, and provide access to data and information. Information is captured throughout the decision guide through statements that classify and document decisions made at every step.

Much of the data and information used as inputs for regulatory modelling comes from external sources. Consequently, understanding how information flows from source to use is essential to modelling. Partnerships are the simplest type of interaction for transferring content between a single provider and user. There are normally agreements detailing the roles and responsibilities of both parties, intellectual property rights, and how data exchanges are coordinated.

For groups, associations, or networks, data and information are captured from multiple sources and processed by a single end user. Group members are independent but operate under agreements that describe authorities, roles, responsibilities and operating procedures for all members. Networks integrate diverse content among many, disparate providers and users to create a holistic view of a multi-faceted system. Network participants are fully autonomous and operate through network rules of conduct.

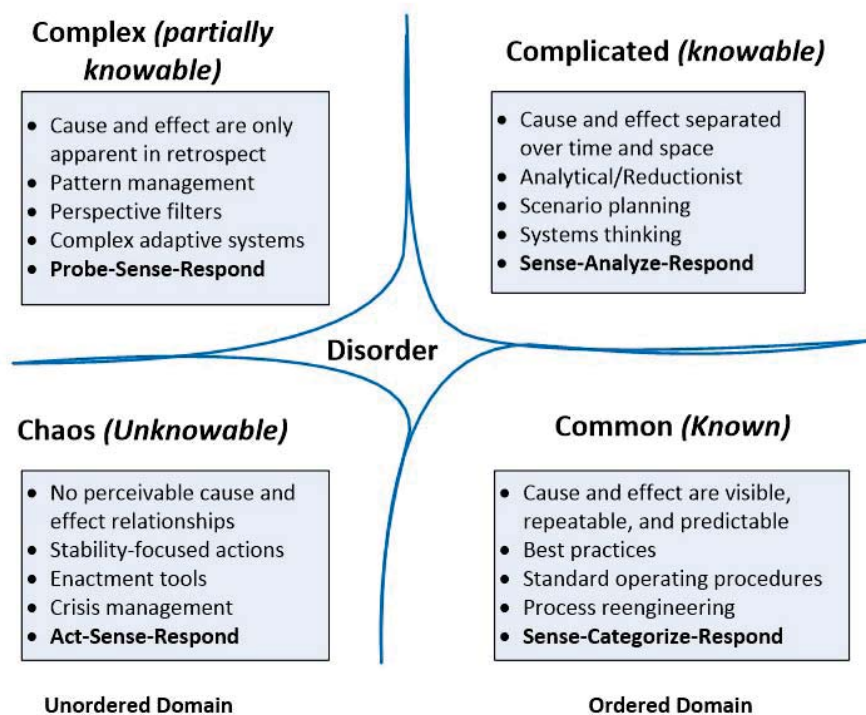
Information markets facilitate information exchanges between many diverse providers and users. Providers may provide repository services or links to repositories, facilitate search and retrieval, convert media and formats to those needed by users, and disseminate information – at no cost, for profit, or something in between. Users search for and acquire information and apply it to addressing issues or solving problems. Information brokers, such as libraries and information centers facilitate the exchange of information between providers and users.

Sequential information markets interpose multiple exchanges and transformations between the source of information and its end use. This can result in processing delays, modification of content, and disconnects at each link in the chain. Despite the inherent weakness of sequential information markets, it is a common form of information flow.

### 3.3 Knowledge

Knowledge is understanding cause-and-effect relationships arising from analysis or synthesis of data or information that facilitates explanation and prediction of physical, natural, or social phenomena. Knowledge may be explicit (codified in reproducible media), or tacit (held in the mind of an individual). Knowledge categorization is considered in the design stage of the decision guide and knowledge to support a model is considered in the knowledge stage of both supply and demand phases of the decision guide.

The Cynefin sense-making framework developed by Kurtz and Snowden, (2003) classifies knowledge into four domains: common, complicated, complex, and chaotic (Figure 1). The domains identify situations where modelling is most suitable and categorize models appropriate to each region. It is critical that a situation be correctly classified in the correct domain and that modelling approaches are appropriate to that domain to minimize the risk of finding an elegant solution to an irrelevant problem.



**Figure 1.** Cynefin Sense-Making / Classification Framework

- *Common knowledge* is generally well known. Guidelines are used to classify work and then rules and standards are followed to do the work. This is the realm of bureaucracy and standards. Work flow and process models are best suited to situations involving common knowledge.
- *Complicated knowledge* is the professional knowledge of experts and consultants. The objective is to provide knowledge-based services in fields such as medicine, engineering, and the law, using existing standards interpreted through explicit knowledge. Some aspects of managing food safety require complicated knowledge. Because the knowledge is well-defined, complicated models tend to be relatively definitive with limited uncertainty. They could play a substantial role in supporting risk management decisions and actions, as well as producing improved outputs.
- *Complex knowledge* may not be explicitly known but is ultimately knowable. The challenge is to discover patterns and understand process in fields such as science and enterprise leadership, using tacit

knowledge, experience, and judgment. Some aspect of managing food safety require complex knowledge that is incompletely- or poorly-defined. Models can provide only a portion of the information needed for decision making. All decisions and actions involving complex knowledge involve significant uncertainty

- *Chaos is unknowable.* There are no discernable patterns or cause-and-effect relationships. Some aspects of regulation, such as the occurrence of a hazardous event are unknowable in advance. Here, we are limited to scenario analysis, sensing the environment, and responding to unpredictable events. Regardless of a models contribution to an overall decision, the resulting outcome will be improved by that contribution.

A four-quadrant framework is also used to describe the usefulness of knowledge, based on the reality and awareness of what we know and don't know (Figure 2). It is important to understand what is known and not known about a situation to avoid discovering a mismatch between model outputs and reality after the fact.

<b>Awareness</b>	<b>Know</b>	<b>Uncertain <math>P &gt; 0</math></b> <ul style="list-style-type: none"> <li>• Assess Risk</li> <li>• Research</li> <li>• Learn</li> <li>• Prepare</li> </ul>	<b>Productive <math>P &gt; 0</math></b> <ul style="list-style-type: none"> <li>• Share</li> <li>• Mobilize</li> <li>• Integrate</li> <li>• Use</li> </ul>
	<b>Don't know</b>	<b>Ignorant <math>P &gt; 0</math></b> <ul style="list-style-type: none"> <li>• Scenarios</li> <li>• Sensitivity</li> <li>• Monitor</li> <li>• Plan</li> </ul>	<b>Wasted <math>P &gt; 0</math></b> <ul style="list-style-type: none"> <li>• Unavailable</li> <li>• Unusable</li> <li>• Duplicated</li> <li>• Lost</li> </ul>
$\sum P = 1$		<b>Don't know</b>	<b>Know</b>
		<b>Reality</b>	

**Figure 2.** Knowing and Not Knowing

- *Ignorance* involves not knowing about a situation and being unaware of what isn't known. We can only estimate the potential impacts of possible events. Ignorance results in surprise that has resulted in major negative intervention outcomes.
- *Uncertainty* involves not knowing about a situation and being aware of what isn't known. Uncertainty sets the stage for research if the situation permits. Uncertainty must be considered in model development, included in model outputs, and referenced in advice stemming from those outputs.
- *Productive knowledge* involves knowing about a situation and being aware of what is known. Because a model is only as good as the knowledge that supports it, we must determine the extent of our knowledge about a situation before developing a model or using model outputs.
- *Wasted knowledge* involves knowing about a situation and loosing or hoarding the knowledge. Wasted knowledge must be recreated or reacquired every time it is needed.

#### 4. Decision Guide

The decision guide transforms the underlying concepts of modelling and content management into concrete actions. A systems approach is used to describe the work flow throughout the modelling process. The modelling principles have been incorporated into the guide structure and the framework is based on the desired attributes described previously. Data and knowledge are considered in designated sections of the guide while information is captured and documented at every step.

The decision guide relates the modelling process to regulatory activities associated with food safety. However, most of the guide is generic. Domain-specific needs comprise only a small part of the overall guide and by substituting requirements for other domains, the guide could be applied to a broad range of complex

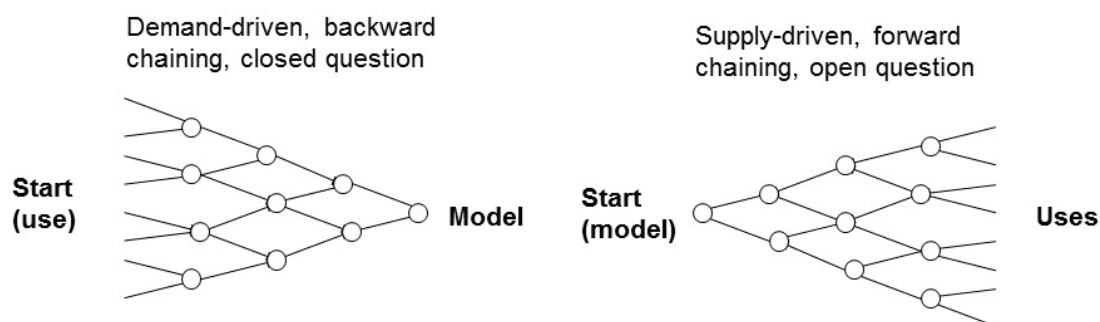
situations. In providing a framework for decision making, the guide helps to structure experience and judgement, but it does not replace them.

The decision guide has four primary uses.

- Help managers determine the appropriateness of modelling for addressing a situation;
- Ensure that model development is targeted to an intended use;
- Evaluate the applicability of an existing model to a range of uses; and
- Demonstrate due diligence and document modelling decisions to withstand challenges.

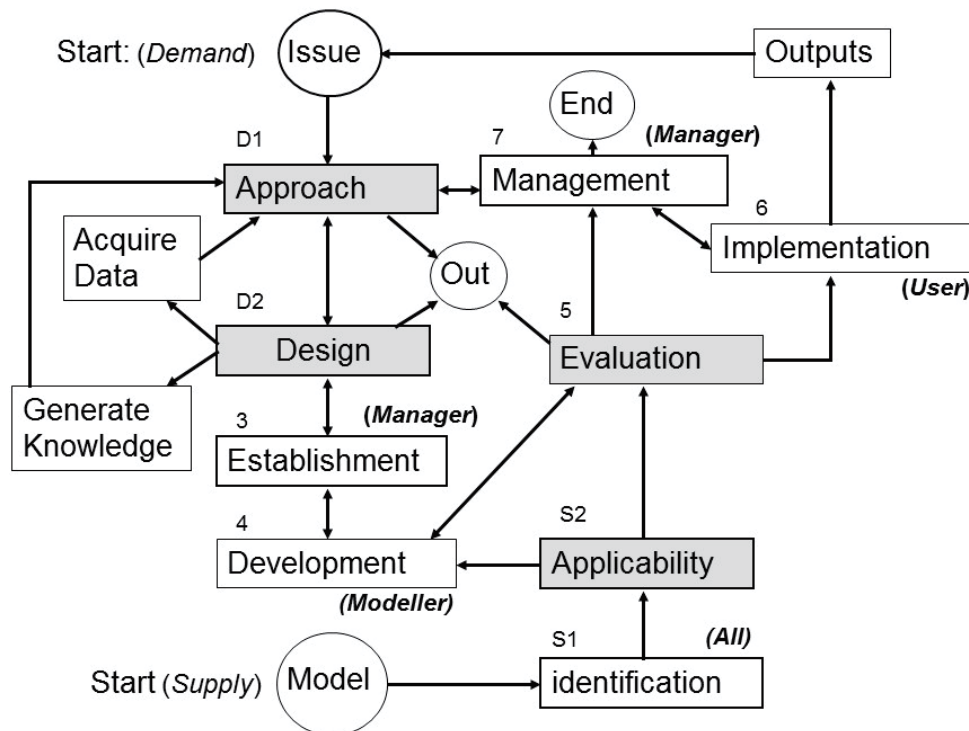
The modelling framework can be used for situations that are driven by an issue (problem in search of a solution) or an existing model (solution in search of a problem). Thus, the framework provides a infrastructure for a “modelling market.” That is, it provides a common language and processes to facilitate interactions among those who develop models (supply), those who use models (demand), as well as those who manage the process.

The strategy for supply or demand approaches to modelling are not mirror images of each other. Demand starts by determining design specifications – a backward-chaining, closed-question that starts with many possible approaches and ends with a specific model (Figure 3). The most effective strategy is to eliminate as many inappropriate possibilities as quickly as possible. In contrast, supply starts by determining applicability – a forward-chaining, open question that starts with an existing model and ends with a set of possible uses. The most effective strategy is to eliminate the fewest potential possibilities as slowly as possible, to arrive at a range of appropriate uses.



**Figure 3** Demand-Based vs. Supply-Based Modelling

The modelling framework shown in Figure 4 consists of four phases that are divided into nine primary stages: demand (approach, design), supply (identification, applicability), project (establishment, development, evaluation), and application (implementation, management). There are also two secondary stages – acquire data and generate knowledge that may be necessary. Overall, there are opportunities for feedback and multiple iterations at almost every stage throughout the modelling process. Clearly, modelling it is anything but a simple, sequential, linear, flow-through process as it is often portrayed in the literature.



**Figure 4. Modelling Framework to Support Regulation**

Each stage consists of a number of steps which are listed here but not described. They are described in detail by Simard (2009). Each step begins with an explanation or description of the topic under consideration. The user is asked to classify a situation or model, using a list of categories that is provided. The user then completes a descriptive statement about the topic, using a template that is provided. Finally, the user decides where to go next in the guide. Category lists and statement templates substantially reduce the time and of effort needed to complete the process. The statements not only provide inputs for a model inventory, they also document the decision process for subsequent justification and demonstration of due diligence.

The guide reflects three different modelling perspectives, those of modellers, managers, and users. Four of the primary stages (identified in Figure 4) are primarily of interest to one of the three perspectives. Four of the primary stages (shaded in Figure 4) promote and facilitate interaction among the three groups: approach, design, applicability, and evaluation. Output from these four stages should reflect a consensus among the three groups to ensure that an appropriate model is developed or selected, it is delivered on time and within budget, and it provides outputs that support desired outcomes. One stage (identification) can be done by any group without interacting with other groups.

### 4.1 Demand

The modelling process may begin by identifying a problem, opportunity, or issue – a problem in search of a solution. This is a “demand” approach, since it originates from a want or need related to a specific regulatory situation. Demand consists of two primary stages – Approach and Design as well as three secondary stages – Knowledge Evaluation, Data Availability, and Data acquisition. At each stage, the guide facilitates decisions to continue, continue with conditions, or exit the guide. Approach and Design involve feedback in that more detail or better information from a latter step may result in re-evaluating a decision made in an earlier step.

*Approach* determines the need for and appropriateness of modelling to help resolve a problem or address an issue. It begins by searching for previously-developed models. Approach also considers data availability and existing knowledge as key precursors to model development. There are three possible results from the

Approach stage: 1) there is an existing model in the inventory and it is implemented; 2) modelling is appropriate and continue working through the guide; or 3) modelling is not appropriate and exit the guide.

*Design* develops model specifications that will be needed to support the intended use. This stage involves interaction, dialogue, and exchange of views among model developers, managers, and users to develop a consensus on necessary model attributes, modelling methods, and design specifications. There are four possible results from design: 1) model specifications are developed and continue working through the guide; 2) knowledge is inadequate to develop a model and create the needed knowledge; 3) available data are inadequate to satisfy use requirements and acquire data; or 4) a model cannot be designed to satisfy the intended use and exit the guide.

*Knowledge Evaluation* – A model is only as good as the knowledge that supports it. Knowledge evaluation examines what we know about a situation, what we need to know to model it, and evaluates whether or not we know enough. Alternatively, it assesses how much of a situation can be modeled, given what is known. Knowledge evaluation often involves successive compromises between what is needed and what can be done. The applicability of similar knowledge to a current situation should be adequately demonstrated before proceeding. Knowledge evaluation involves four sequential steps: knowledge needs, existing knowledge, knowledge gap, and create new knowledge.

*Data Availability* – Models without data are academic exercises. Data availability determines whether or not the data needed for running a model are available when, where, and how they are needed. Data availability has two perspectives – proposed or existing models. The former considers data that are available to support model development. The latter focusses on data that are needed to run an existing model. Determining data availability may involve feedback as substitutes or surrogates must often be used instead of preferred data. Determining data availability involves four steps: needs, attributes, accessibility, and processing.

*Data Acquisition* - If the data needed to run a model are not available, it may require months or years of effort to acquire it. Consequently, a decision to proceed must reflect the available time, cost, and effort. Data acquisition and information system development may proceed in parallel with model development or in advance of model development. Data acquisition involves three steps: recurrence, data collection, and data management.

## **4.2 Supply**

The modelling process may begin with an existing model with potential applicability to one or more problems, opportunities, or issues – a solution in search of a problem. This is a supply approach in that it originates from an existing model that someone proposes to implement. A supply approach begins with existing specifications and determines the scope of applicability of the model. This not only reverses the demand approach, but also shifts decision strategies from eliminating options as soon as possible to retaining options as long as possible. The supply approach consists of two stages – Identify and Applicability

*Identify* searches for and describes existing models that could be used to address problems and issues of interest to a regulatory agency. Results from Identification are transferred to Applicability.

*Applicability* reverses the process used in approach and design. That is, it starts with existing design specifications and determines the scope of appropriate uses. There are four steps in applicability: model specifications, data availability, supporting knowledge, and suitability. Results from Applicability are transferred either to Evaluation for approval or to Development for enhancement and/or adaptation.

## **4.3 Project**

The project phase shifts the modelling process from decision making to doing. Although this phase typically involves step-by-step progression, some aspects of the process are iterative. There are three stages in the project phase: Establishment, Development, and Evaluation.

*Establishment* constitutes a modelling project as an accountable organizational mandate to use allocated resources for developing a model to achieve specified objectives. A modelling project uses the modelling process to accomplish its objectives. Project establishment is the essential first step of project management.



Project establishment provides an organizational structure for modelling. Establishing a modelling project involves three considerations: project mandate, project infrastructure, and project planning.

*Development* is the technical heart of the framework. It involves constructing and testing a model. This is the stage in which analytics plays a dominant role. Because some sort of model can always be developed, Development always transfers results to Evaluation. This is a two-way flow, however, as a model may be sent back to development for additional work. Whereas the other four project stages should require only a few days to a few weeks to complete, development may require a few weeks to a few years. Development includes three steps: conceptualization, construction, and verification.

*Evaluation* uses criteria and indicators of the quality, authoritativeness, and acceptance of a model and its outputs to determine approved uses. Because modelling typically requires compromise as it progresses, it is important to determine the extent to which a finished model meets the initial specifications and satisfies use requirements. This leads to formal approval for specified uses and documentation in the model inventory. Evaluation consists of three steps: validation, endorsement, and approval.

#### **4.4 Application**

Application is the phase in which the (sometimes substantial) time, effort, and resources required to develop a regulatory model provide a return on that investment. If a model isn't used for its intended purpose, it represents a cost with no benefit. Although modelling results in learning and creating new knowledge, these are primarily academic benefits which support but are not central to the core business of a regulatory agency. Consequently, the Design or Applicability stages are important to maximize the likelihood that once developed or selected, a model will be used for its intended purpose and its use will generate desired outcomes. There are two stages to Application – Implementation and Management.

*Implementation* involves conducting analyses to accomplish the intended purpose of the model, interpreting outputs, providing results to the intended recipients, and using the results to support decision-making. This stage may be bypassed if the final model is deemed inappropriate for the intended use. Alternatively, the regulatory agency could compile an inventory of approved models for future use. Implementation is a three-step process: analysis, communication, and execution.

*Model Management* involves capturing the model and its attributes in a model inventory, describing its development history, documenting the model and its implementation, summarizing lessons learned, and making it available for future use. This completes one cycle of the modelling process. Model management comprises four steps: documentation, project evaluation, case study, and inventory.

### **5. Conclusion**

The scale and complexity of many regulatory situations precludes directly linking intervention alternatives with desired outcomes. By simplifying reality, modelling provides a method for holistically analysing the many components of a complex system. However, modelling is part art and part science, which provides a basis for legal challenges to model outputs to counter the negative impacts of complying with regulatory decisions. The modelling framework described here not only provides managers with a flexible structure for efficiently and effectively managing the modelling process but also demonstrates due diligence in model development.

The proposed framework has a strong conceptual foundation based on a systems approach, best modelling practices, and good framework design. It also incorporates the content (data, information, and knowledge) that define and describe a model. A management decision guide emphasizes appropriateness for an intended use, supports a supply (existing model) or demand (new model) approach to modelling, facilitates interaction among modellers, managers, and users, and documents due diligence when facing challenges. The framework comprises four phases – demand, supply, project, and application. Each phase consists of two to four stages which, in turn, include a number of steps and considerations. Each step involves an explanation, classifying a situation, writing a brief statement, and selecting the next step.

Using the framework increases the likelihood that a model will be appropriate for its intended use and provide the right answers to the right questions, in time to make the right decision. It eliminates wasted time, expense, and effort of developing a model that does not adequately represent a situation. It also reduces the

chance of misusing an existing model that does not represent a situation. Finally, it documents the modelling process to enable a model and its outputs to withstand legal challenges. Although the framework cannot replace experience and good judgement, it provides an easy-to-use structure for organizing and managing the complexities of modelling in a regulatory environment.

## References

- Duram, Sharon (2006) Food Safety in the 21<sup>st</sup> Century. USDA, Agricultural Research Magazine, Vol. 54 No. 10  
<http://www.ars.usda.gov/is/AR/archive/oct06/index.htm>
- Frazil, A., Sanchez, R., and McEwen, S. (2008) Choices, Choices: The Application of Multi-Criteria Decision Analysis to a Food Safety Decision Making Problem, Journal of Food Protection, No. 11 p. 2168-2373
- Garner, G., Dube, C., Stevenson, M., Sanson, R., Estrada, C., and Griffin, J. (2007) Evaluating Alternative Approaches to Managing Animal Disease Outbreaks - The Role of Modelling in Policy Formulation, Veterinaria Italiana Vol. 43, No. 2, p. 285.
- Greenberg, M and L. Cox (eds.) (2012) The Economics of Risk Analysis, Society for Risk Analysis Jour. Of Risk Analysis Vol. 32 No. 11
- Holmes, J., Graham, J., Mckone, T., and Whipple, C. (2009) Regulatory Models and the Environment; Practice, Pitfalls, and Prospects, Risk Analysis, Vol. 29, No. 2, p.159-170.
- Kurtz, C. and Snowden, D. (2003) The New Dynamics of Strategy: Sense-Making in a Complex and Complicated World, IBM Systems Journal, Vol. 42, No. 3.
- National Academy of Sciences (2007) Models in Environmental Regulatory Decision Making, National Research Council, Committee on Models in the Regulatory Decision Process, National Academies Press, Washington DC, 267 p.
- Simard, A. (2009) Canadian Food Inspection Agency, Modelling Framework, unpublished report, July 15, 2009, 257 p.
- United Nations, World Health Organization and the Food and Agricultural Organization (2006) Codex Alimentarius: Understanding the Codex, Rome, IT 39p. <http://www.codexalimentarius.org/about-codex/understanding-codex/en/>



# Online Community Projects in Lithuania: Cyber Security Perspective

**Aelita Skaržauskienė, Darius Šttilis, Monika Mačiulienė, Gintarė Paražinskaitė and Agnė Tvaronavičienė**

**Mykolas Romeris University, Vilnius, Lithuania**

[aelita@mruni.eu](mailto:aelita@mruni.eu)

[sttilis@mruni.eu](mailto:sttilis@mruni.eu)

[giparaz@mruni.lt](mailto:giparaz@mruni.lt)

[a.tvaronaviciene@mruni.eu](mailto:a.tvaronaviciene@mruni.eu)

[maciuliene@mruni.eu](mailto:maciuliene@mruni.eu)

**Abstract:** Within the context of smart and inclusive society cyber security is an important issue, which must be analyzed and discussed in the field of science and practice. Collective intelligence, which emerges in the activities of online communities, is a new quality of civil engagement that grants more effective decisions and compliance with societal needs. Various social technologies created possibilities for society members to communicate despite the limitations of the physical world but they have brought high prospects for sophisticated crimes and other violations of rights and obligations of users, administrators and states as regulatory bodies. This article connects several independent fields of research: analytics, social technologies, civil engagement, collective intelligence and cyber security in order to reveal the main threats of using social technologies during the process of engaging society into socially responsible activities. The sources of data are growing and data mining could be used in variety of ways. Analytics allow to combine different observations in order to see new patterns. Risks related to personal data processing in the virtual community networks are discussed together with analysis of related EU legal regulations on personal data protection. The authors stress importance of personal data protection in online networks and legal problems that arise in networked society. As a result, the main trends were identified in the context of online community projects in Lithuania based on quantitative public opinion survey conducted in 2013.

**Keywords:** online community, social networks, collective intelligence, legal issues, civic engagement, data protection.

---

## 1. Introduction

Following the Internet expansion, organizations and movements have evolved from bureaucratic/centralized to both decentralized and distributed networks. This evolving change towards de-centralization and democratization has started to impact business, governments and society at large (Malone, 2010). "Since the future is basically unpredictable and uncertain, society must rely on creative initiatives from the citizens to be able to create the desired future" (Johannessen, 2001). Cyber security issues, within the context of smart and inclusive society, are important aspects, which must be analyzed and discussed in the field of science and practice. Collective intelligence, which emerges in the activities of online communities provides a new quality of civil engagement that grants more effectiveness and compliance with societal needs. Various social technologies have created possibilities for society members to communicate despite the limitations of the physical world but they have brought high prospects for more sophisticated crimes and other violations of rights and obligations of users, administrators and states, as regulatory bodies. This article connects several independent fields of research: analytics, social technologies, civil engagement, collective intelligence and cyber security in order to reveal the main threats of using social technologies during the process of engaging society into socially oriented activities. The sources of data are growing and data mining could be used for different reasons. Analytics allow mixing different observations together in order to see new patterns. The risks related to processing of personal data in the online community networks, also related EU legal regulation on personal data protection are discussed.

The article focuses on the legal issues of cyber security identified in the empirical research on the involvement process of online communities in different civil activities. The authors stress importance of personal data protection in online networks and legal problems that arise in networked society. As a result the main trends in the context of online community projects in Lithuania were identified based on quantitative public opinion survey conducted in 2013.

## 2. Regulatory framework for online communities

Online communities provide useful tools for communication and information exchange, however online networks and the use of such networks raise many questions regarding users' data protection. Typical

personal data published by users include user name, sex, birthday, age, and contact information (e.g. e-mail address, telephone number, address), and instant messenger screen name. Depending on the community site, users may also be able to post additional information such as their sexual orientation, where they work or attend school and their religious and political affiliation (Henson et al, 2011). The core of online communities' sites is composed of users' profiles showing an expressed list of user connections and relations. The basic idea is that members will use their online profiles in order to become part of an online community of people with common interests (Bygrave, 2012).

Personal information, which is very sensitive, and is processed in the context of online networks, may be used or transferred for illegal purposes. Personal data published on social network sites can be used by third parties for variety of purposes, including commercial, and may pose major risks such as identity theft, financial loss, loss of business or employment opportunities and physical harm (Opinion 5, 2009). *The Article 29 Working Party* is a platform for cooperation, composed of representatives of the EU member states national data protection authorities and the European Commission. This organization identifies several legal problems related to data protection in social online networks: The problem of security and default privacy settings; Information to be provided by virtual-social network provider; The problem of sensitive data; Processing of data of non-members; Third party access; Legal ground for direct marketing; Retention of data; Right of users etc. Also, there is the problem of jurisdiction i.e. whether EU has jurisdiction over activities of international players, originated from non-EU countries.

What is the regulatory framework for data protection in online communities? In 1950, a fundamental right to "respect for private life" was included in the European Convention for the Protection of Human Rights and Fundamental Freedoms. According to the article 8 of the convention, everyone has the right to respect for his private and family life, his home and his correspondence (Convention for the Protection of Human, 1950). On the global scale, main international legal act concerning data protection is Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data of 1981. The purpose of this convention is to secure for every individual, whatever his nationality or residence, respect for his rights and fundamental freedoms, and in particular his right to privacy, with regard to automatic processing of personal data relating to him in the territory of each Party (Convention for the Protection of Individuals..., 1981). The convention consists of main principles, which can be applied in case of online social networks (for example, principle of data quality or principle of data security). However, these principles do not function properly in the light online social networks (mainly because of abstract character of these principles) (Šttilis, Gutauskas, Malinauskaite, 2012).

In EU context, there are two directives regarding data protection: General data protection directive No. 95/46/EC and Directive on privacy and electronic communications No. 2002/58/EC. The provisions of General data protection directive apply for social networks providers in most cases. However, the directive was created when there were no social networks in cyberspace at all. Therefore, the main principles and provisions of the directive cannot solve data protection issues of online nature and most of the issues remain uncovered (Legal and privacy..., 2012). For example, it is unclear who controls data: owner/manager of online network, application provider, or user of such network. Because of such unresolved issues. Directive 2002/58 on Privacy and Electronic Communications, otherwise known as E-Privacy Directive, is an EU directive on data protection and privacy in the digital age. It deals with the regulation of a number of important issues such as confidentiality of information, treatment of traffic data, spam and cookies. The Directive complements the General Data Protection Directive and applies to all matters that are not specifically covered by that Directive. We may note that not every privacy issue in online networks is related to electronic communications. However, the role of communications in online networks should not be undermined and this directive should govern issues related to it.

European Commission decided to revise the data protection legislation and approached to draft and prepare new regulations for data protection. Viviane Reding, current EU Justice Commissioner, in a speech held on a Conference in London 2011, pointed out that a need for a more comprehensive and coherent approach in the EU policy for the fundamental right to personal data protection (Reding, 2011). However, the regulation is under discussions in EU institutions until now.

Other notable international efforts towards data protection viable legislation include pioneering standard-setting report published by UNESCO on Internet freedom titled "Freedom of Connection – Freedom of

Expression: The Changing Legal and Regulatory Ecology Shaping the Internet” and summary of ten core principles ranging from universality and equality, accessibility, and rights and social justice to diversity and network equality published by Access (also known as Access Now.org), an international Internet advocacy group dedicated to an open and free Internet.

Actions discussed in this section clearly emphasize that attention must be paid towards cyber security issues in order to develop safe and reliable environment for people, who wish to engage and wish to generate ideas for greater welfare of society. The main challenge now is to exploit the potential of new media while not compromising civil liberties, including the right to freedom of expression, to education and to privacy.

### **3. Internet user’s analytics in Lithuania**

In Lithuania and other Eastern and Central European countries of post-soviet bloc, civic engagement of society is low when compared with Western European partners. This problem is not new and often appears in the horizon of scientists, who research society using different perspectives. One of the goals of scientific project currently carried out by our scientist group at Mykolas Romeris University (Lithuania) is to reveal the legal problems, which might limit civic engagement via networks, as well as create obstacles for the emergence of collective intelligence as more effective intellectual instrument for overcoming social challenges. Quantitative research on the extent and current trends and of the society’s engagement and participation in building collective intelligence was carried out as part of this project. Having in mind the fact, that such difficult and multi-dimensional questions can be resolved only after evaluation of existing situation and factors influencing it, we chose to employ analytics of quantitative data as research method. Analytics is a process oriented towards disclosure of most important determinant factors and it was used to answer questions below:

- Who are the frequent Internet users in Lithuania?
- What activities people usually carry on in Internet?
- What main problems people using Internet are facing?

All three questions will be analyzed on two levels. Firstly, we are going to provide a **Profile of frequent Internet** user together with the **Set of activities** considered to be most popular on the web. Answers to the third question will empower us to list problems needed to be taken into account, while designing perspectives of more inclusive society. Finally, we will present the Profile of internet user involved in socially oriented activities on the web, problems and barriers he/she faces while using online tools. This profile will allow us to proceed further with the research and prepare grounds for a qualitative research stage in terms of identification of respondents for in depth interviews.

The selection of survey respondents were undertaken by respecting general rules of a random stratification sample and the specifics related to the participation in the process of collective intelligence emergence. Sample (N=1022) included 478 males and 544 females aged 15-74 in all districts (urban and rural areas) of Lithuania, which guarantees a statistically reliable representation (with the confidence level of 95%) of the Lithuanian population. Public opinion survey was carried out using the method of direct interview at respondents’ houses using computerized and standardized questionnaires. Interviewed respondents represented the overall Lithuanian population by the major socio-demographic characteristics (using stratified random sampling). After collecting the survey data, statistical study was carried out using SPSS for Windows. Statistical relationships between attributes were calculated by using a chi-square ( $\chi^2$ ) tests. Significance level of  $p < 0.05$  was chosen to calculate statistical reliability.

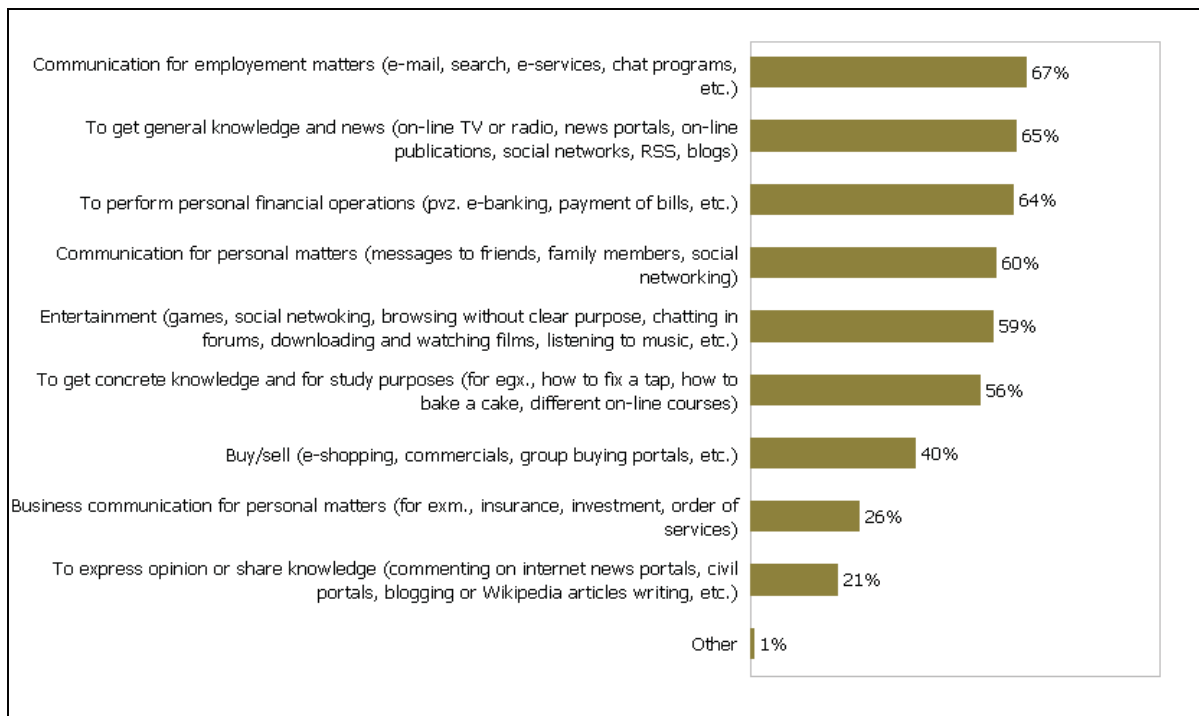
Further, analysis of public opinion survey will be presented. Main attention will be drawn towards composing Profile of the user based on legal risks related to deeper involvement into activities of socially oriented platforms. According to results of public opinion survey, all respondent can be divided in 6 groups, but only three groups gathered more than 10% of respondents. These groups are represented in Table 1.

**Table 1:** Frequency of using Internet

Frequency	Results
Every day	44%
Few times a week	14%
Once a week	5%
Few times per month	2%
One per month and less, but minimum once per 3 month	1%
Less than once per 3 months / Not using at all	34%

We can see, that 44% of respondents use internet every day, and only 34% use Internet infrequently (less than one time per three months) or do not use it at all. Remaining 22% use the Internet less than few times per week. It can be noted, that most frequent Internet users are younger (39 years old or less), well-educated people living in 3 biggest cities of the country ( $p < 0.05$ ). Respondents, who do not use Internet or use it very rarely, usually are 50-74 years old, living in rural areas of country and have lower income ( $p < 0.05$ ).

In general, most popular activities online appear to be communication related to professional aspects of life (67% respondent have chosen this option). Activities related to generation of general information and knowledge are close second (65%). However, Internet users who use Internet more than few times a week are not always active in visiting socially oriented websites ( $p < 0.05$ ). As we will discuss later, socially oriented activities are not listed among the most popular activities online. Figure 1 below shows what activities are most popular online in Lithuania.



**Figure 1:** Most frequent activities in Internet

Only 21% of respondents answered that they share their opinion or knowledge online (comments in various websites, community forums, blogs, etc.). This reveals that Lithuanian people are not inclined to participate in socially oriented activities neither offline nor online. These findings encourage to further research of low civic engagement.

Personal communication using online channels was indicated as one of the most popular options by respondents. 61% of respondents using Internet are using various websites of online communities and social networks and only 33% of respondents are not registered to such activities. 6% of respondents said that they have accounts but do not use them. Most active users of social networks and online communities are people

aged 15-29 ( $p < 0.05$ ). Most popular and best-known social network in Lithuania is Facebook (82% of respondents, using Internet, mentioned it). Different situation can be observed, when respondents were asked to name socially oriented networks operating in Lithuania. Only one well-advertised network [www.darom.lt](http://www.darom.lt) gathered 41% of respondents. Other projects were mentioned by less than 20% of respondents. Deeper analysis shows, that respondent mostly use social networks and online communities in order to pursue personal interests related to hobbies or other areas of personal life (74%). Thus, socially oriented platforms were not mentioned often.

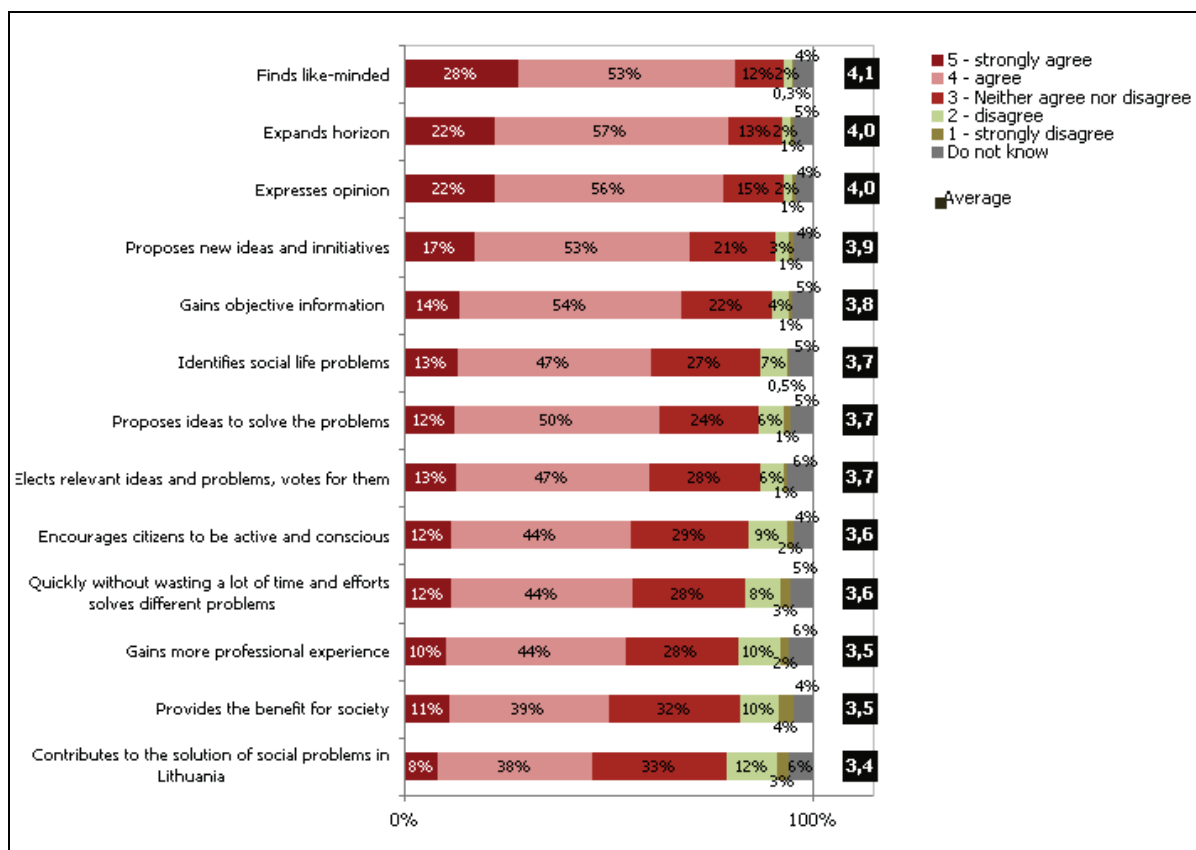
Table 2 below shows number of survey respondents interested in visiting platforms tackling societal problems as compared to general online communities and social networks.

**Table 2:** Visiting websites of online communities and social networks versus visiting websites oriented towards social problems solving

Option	Result
Uses websites of online communities/social networks (not oriented towards solving of social problems)	54%
Uses websites of online communities/social networks (also and oriented towards solving of social problems)	7%
Uses websites oriented towards solving social problems (but not websites of online communities/social networks)	0,4%
Does not use websites of online communities/social networks or other website oriented towards social problems solving	38%

The number of people involved in online communities and social networks not oriented towards solving of social problems is 54% of all respondents using Internet. These high numbers show, that respondents are actively communicating using social technologies. However, number of respondents involved in activities of socially oriented virtual communities and social networks is only 7%. Obviously, people do not involve themselves in socially oriented activities even if they are organized in familiar online environments. Survey also revealed that more educated people use social networks more frequently ( $p < 0.05$ ). It should be noted that people, who are not members of social networks and online communities at all (38%), are usually 40 or older, have only professional education, are married or living with partner ( $p < 0.05$ ). In order to proceed with analysis of quantitative data and for the sake of explicitness, the users were divided into three groups based on their activity in socially oriented online projects. "Strong users" identified as respondents, who visit socially oriented platforms every day. "Medium users" are involved in such activities few times a week. "Weak users" join such socially oriented networks once a week or less. It is obvious that "strong users" are active in most activities online. That leads to a presumption, that Internet itself is inclusive which means that people who are acquainted with this technology start to use it more and extensively. "Weak users" of Internet are very passive considering their involvement in socially oriented activities online.

Having in mind information presented in Figure 1, comparative analysis of public evaluation of possible applications of online communication will be conducted.



**Figure 2:** How do you evaluate the different aspects of online communication?

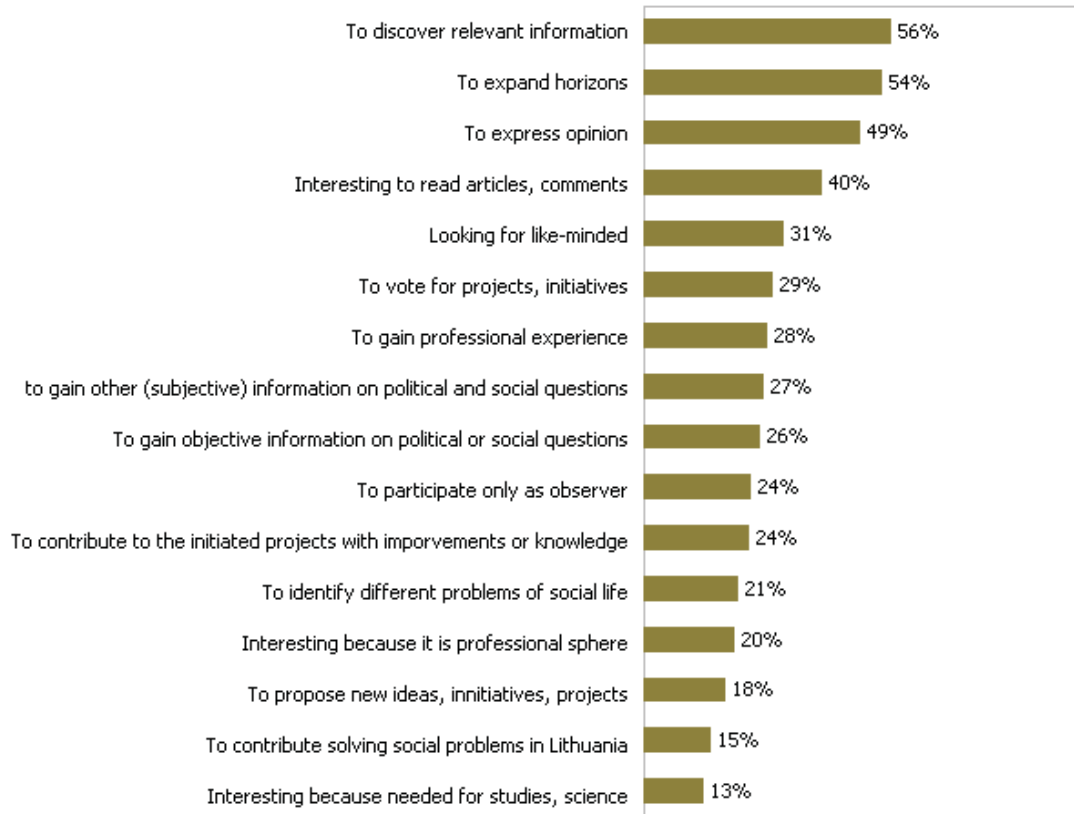
Figure 2 reveals that respondents know and understand the perspectives and benefits of Internet based social involvement. For example, high positions were granted for such active behavior as finding like-minded people, expressing opinion or proposing of new ideas. Unfortunately, people do not indicate that they involve themselves in such activities (see Figure 1). They rank all possible outcomes more positively than negatively, but show no interest into realization of such expectations.

Respondents, who are using websites with socially oriented goals, also were asked to identify what particular activities they perform there, results are presented in Figure 3.

Data presented above shows, that most popular activities are quite passive and related only to observation of processes happening in websites, oriented towards social problems solving. For example, getting relevant information (56%), broadening of views (54%), getting acquainted with interesting information and comments (49%), searching for like-minded people (31%), getting more professional experience (28%) to compare with such active activities as expressing of own opinion (49%), voting for projects or ideas (29%), improving of projects by using own knowledge and skills (only 24%), suggesting new ideas or projects (only 18%) and giving own input for social problems solving (only 15%). Such distinction shows that even those respondents, who are using websites, oriented towards tackling of social problems, mostly are inactive and mainly susceptible to observing the ongoing processes rather than taking part in them.

#### 4. The cyber security perspective in online communication

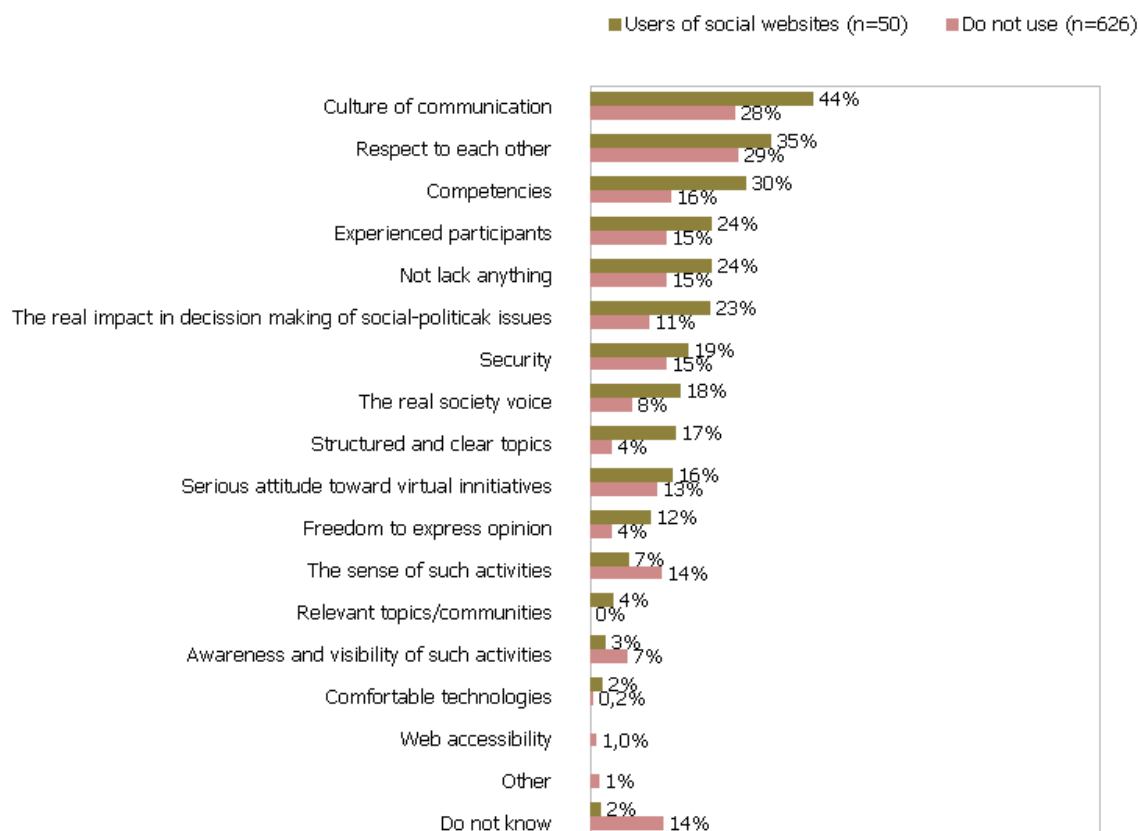
Authors in article „Cyber Security and Civil Engagement: Case of Lithuanian Virtual Community Projects” (Skaržauskienė et al, 2014) analyzed the Internet accessibility in Lithuania. A paradox have been observed while analyzing usage of social technologies in Lithuania – residents enjoy using technologies for work, leisure and personal every day needs, but most of them are inactive users of various socially oriented platforms. While more than 75% of residents (aged 16-74) have access to Internet, only 4-12% of population are involved in socially oriented activities. This paradox encourages further investigation of complex question: what online platforms lack in order to attract more users in socially oriented activities online?



**Figure 3:** Socially oriented activities on websites

Respondents of public opinion survey also were inquired why they do not use online communities and social networks in general. Most popular answer was the 'unacceptable culture of such communication method'. Other popular answers include 'lack of time' and 'having no interest in such activities'. From legal perspective, 9% of respondent choose 'the lack of privacy' as a reason for limited use of social networks and virtual communities. It is interesting that least popular answers are related to low level of governance feedback to expressed opinion, danger to express opinion, and low security of using it. Also, respondents were asked to list features that socially oriented online platforms. Results presented in the figure below (see Figure 4).





**Figure 4:** What does social websites lack the most?

From the legal perspective only few aspects of the survey can be discussed. First of all, it is the feedback on activities. Only 11% of respondents noticed that one of the weaknesses of social oriented websites is the absence of practical influence of virtual activities on decision-making. Second, the issues of security were mentioned. Only 19% of respondents identified this choice as a weakness, having in mind that websites, oriented toward social problem solving, are not secure. From such information, one conclusion can be drawn: respondents do not consider legal issues of online communication as critically important. Thus in this area one more paradox arises. When respondents were directly asked about the advantages in general of using Internet, they positively evaluated all answers related to socially orientated activities. This statement can also be confirmed by short review of public opinion survey on various aspects of online related to security and regulation (in-depth analysis of these aspects is provided in article „Cyber Security and Civil Engagement: Case of Lithuanian Virtual Community Projects” (Skaržauskienė et al., 2014)). Respondents find almost all listed legal risks equally important and had strong opinions (more than 70%) that strict liability of virtual communities’ members must be envisaged if they violate the rights of other people. Also, it was strictly recommended to think about liability of administrator of networks for the content of networks as well as the need of detail regulation of activities of online communities. Thus, such opinion of respondents should be evaluated only having in mind the previous answers related to identification of main reasons why people do not involve socially oriented Internet projects. If legal aspects were not dominant previously, it is not believable, that legal aspects are so important for respondent who are not involved in socially oriented activities in web and do not find such activity attractive. Likely, people, who are not involved in such activities, cannot identify by themselves independently, what problems they might face.

## 5. Conclusions

The quantitative research results helped us to create the **Profile of frequent Internet** users in Lithuania, where 58 from 100 people are using internet daily or few times per week. Frequent Internet users are younger than 39 year old, well educated, living in the biggest cities of the country. Mostly they use Internet for communication or looking for professional or general information. Frequent Internet users use social networks and online communities in order to realize some personal interests, connected mostly with hobbies or other

areas of personal interests. They perform passive activities such as getting actual information or broadening of view, and mostly escape from active behaviors as commenting or sharing information or knowledge.

Despite of **high accessibility** of Web in Lithuania, people are not inclined to join socially oriented activities. This fact creates an obvious finding that accessibility is the condition but not a **catalyst** for increasing the social involvement of society. Even those respondents, who are used to visit websites, oriented towards social problems solving, mostly are not active and mainly susceptible to observe the ongoing processes rather than take a part in it. People using Internet every day are more often involved in socially oriented activities, and it could be concluded that **digital competencies** in general have a positive influence on online civic engagement.

From the cyber security perspective, respondents do not rank the legal risks as critically important, but they are **aware of cyber security issues** and strongly support most offered ideas about safe and secure operations online. It shows that people in Lithuania still lack experience in online civic activities and cannot identify independently, what problems they might face in virtual space. The united effort is necessary – from the government and law enforcement, to the general public – to meet the evolving challenges in securing cyberspace.

## **Acknowledgments**

The research is funded by European Social Fund under the initiative „Support to Research Activities of Scientists and Other Researcher (Global Grant)“, and administrated by Lithuanian Research Council (grant No. VP1-3.1-ŠMM-07-K-03-030, project title “Influence of Social Technologies on the Development of Collective Intelligence in Networked Society”).

## **References**

- Article 29 Data Protection Working Party (2009) *Opinion 5/2009 on online social Networking*, Brussels, Belgium
- Convention for the Protection of Human Rights and Fundamental Freedoms. Rome, 4.XI.1950. [online]  
<http://conventions.coe.int/Treaty/en/Treaties/Html/005.htm>
- Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Strasbourg, 28.I.1981, [online] <http://conventions.coe.int/Treaty/en/Treaties/Html/108.htm>
- Byggrave, L.A.(2012) Legal and privacy challenges of social networking sites, Doctotal thesis, University of Oslo
- Henson B., Reyns B., Fisher B., (2011) „Security in the 21st Century: Examining the Link Between Online Social Network Activity, Privacy, and Interpersonal Victimization,“ *Criminal Justice Review*, 36(3).
- Johannessen, J. A., Olsen, B., Olaisen, J. (2001) “Mismanagement of tacit knowledge: the importance of tacit knowledge, the danger of information technology, and what to do about it”, *International Journal of Information Management*, Vol. 21, Issue. 1, pp. 3-20.
- Malone, T. W., Laubacher, R. and Dellarocas, C. (2010) “The Collective Intelligence Genome,” *MIT Sloan Management Review*, 51, 3: 21-31.
- Reding V.(2011) Assuring data protection in the age of the internet. London, 20 June 2011, [online]  
[http://europa.eu/rapid/press-release\\_SPEECH-11-452\\_en.htm?locale=en](http://europa.eu/rapid/press-release_SPEECH-11-452_en.htm?locale=en)
- Skaržauskienė, A., Tvaronavičienė, A., Paražinskaitė, G. (2013) „Cyber Security and Civil Engagement: Case of Lithuanian Virtual Community Projects“, International Conference ECCWS, Greece
- Stitilis D., Gutauskas V., Malinauskaite I.(2012) Personal data protection in virtual social networks: legal environment. *Societal Innovations for Global Growth*,1(1).

# Pocket Data Analytics: An Introduction to Practical Applications of Course Analytics in Online Learning

John Vivolo

Polytechnic School of Engineering, New York University, USA

[John.vivolo@nyu.edu](mailto:John.vivolo@nyu.edu)

**Abstract:** The concept of using academic analytics to understand trends in a student population is not a new idea. However, much of the analysis of these massive numbers live in the “Massive or Big Data” analysis of information for whole programs. There is little evidence that instructors are using smaller samplings weekly, or even daily analytics, to improve performance and engagement. For the purposes of this paper, let us call this data: Pocket Data Analytics. This paper will provide an introduction to the practical application of analytics in which instructors can use to proactively improve student engagement and performance. These changes would occur during an on-going semester. According to Swan (2012), “Online learning seems to be on the frontlines of the learning analytics movement, and so online educators need to inform themselves about learning analytics”.

**Keywords:** Pocket data analytics, online learning, learning analytics, practical application, student engagement, student performance

---

## 1. Introduction

Analytic numbers generated in a fully online course appear, at first, as if a storm of numbers without order or purpose. According to Olmos and Corrin (2012), “With the rise of the use of learning analytics to provide ‘actionable intelligence on students’ learning, the challenge is to create visualizations of the data, which are clear and useful to the intended audience.”

With the use of a learning management system’s statistics tool, some semblance of order begins to become clear. As a result, two specific questions come to mind:

Can the non-technical instructor find order in the chaos of the online learning data storm?

Are there practical uses of focused course analytics (Pocket Data Analytics) to improve interaction and performance during a semester?

The result was a term I have come to call, “Pocket Data Analytics”. Rather than looking at large scale data, the purpose of this method is to get instructors to focus on smaller patterns within a single course, during a specific time period, such as a week. The intent is to have a method in which to introduce the concept of academic course analytics as a practical tool and not, as Dringus (2012) states, “a large and unwieldy ‘data dump’ of transaction level data [that] offers little value for the instructor if the yield of data persists in its raw form.”

## 2. Order from Chaos

Before diving into course analytics, let us return to the storm analogy for a moment. The analogy is deceptive, just as storms are deceptive. Storms are not chaos. They appear, at first, to be a random occurrence of wind, rain, and flooding. However a storm, like a hurricane, has a distinct pattern. Hurricanes only happen in certain areas and under certain circumstances. Online learning analytics have a pattern as well. Scientists often struggle to understand this pattern, but so do individuals caught in the storm. Those who see the storm coming apply changes to their environment according to experiences of the past, as well as incorporating suggestions from local experts who have studied these occurrences in detail. According to Dringus (2012), “The data trail, across the online course, may be viewed as the culmination of the ‘lived experience’ of the learner and the instructor. That lived experience of the online course participant – student and instructor – needs to be visualized as a data trail that is traceable and interpretable.”

Regardless of the topic being taught, instructors do not need to be an expert in statistics or mathematics to find a practical use in an online course to improve interaction, performance and engagement in smaller weekly doses.

According to Vernon (2012), “Instructors might be able to launch more customized interventions for at-risk students if they had information showing student performance within specific LMS activities.” Consider this idea of “at-risk” students as the focus of all instructor efforts regardless of online or not. It makes sense for

instructors to begin to utilize all available tools, including course analytics. As per Olmos and Corrin (2012), “Focusing on the most needed insights can motivate creative data collection and representation, even if it involves more work and time.”

### 3. Classic Methods of Improving Performance

A white paper by IBM Software Group (2001), states in part, “One of the most effective tools at hand for schools to improve overall performance is early intervention for outliers; both at-risk students and very high performing students.” When looking at student issues of performance, intervention and interaction, and if we set out to create improvements, we usually have two approaches: (1) reactive – making changes to a course after poor student performance on an assignment/exam; and (2) proactive – setting up preventative measures prior to an assignment/exam.

This paper will, however, focus on at-risk students rather than high performing students.

### 4. Three Types of Analytics for Online Learning

“Some forms of learning analytics provide the ability to generate data reports in real-time, minimizing the time delay between when the data is captured and when actions can be implemented,” (Elsa, 2011). There are three types of analytics captured in most learning management systems that provide real-time reports:

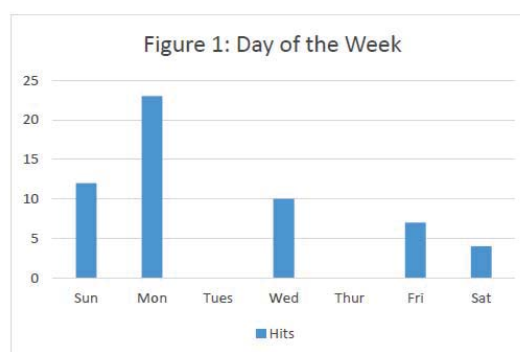
- **Time based:** when students access the course (day of the week, time of the day, etc.)
- **Individual assignments/contents:** how often a student “hits” a single assignment/content.
- **Discussion Board:** how and when are students active in a discussion board forum?

According to Picciano (2012), “Every student entry on a course assessment, discussion board entry, blog entry, or wiki activity could be recorded, generating thousands of transactions per student per course. Furthermore, this data would be collected in real or near real time as it is transacted and then analyzed to suggest courses of action. Analytic software is evolving to assist in this analysis.” The idea of Pocket Data Analytics, however, suggests looking at smaller increments of data for perhaps a week, day, or even for an individual item. While there are many other layers of course analytics, this article will focus on these three types as a recommendation for instructors.

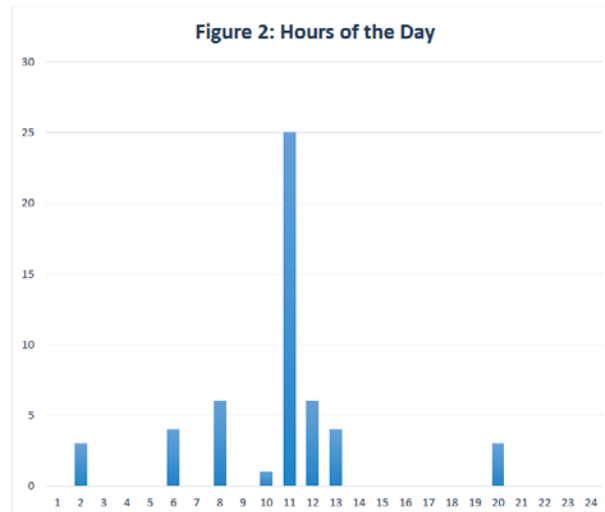
#### 4.1 Time-based activity:

Let us look at two real course examples of time-based activity. Figure 1 represents one week during a fully online, graduate level course. This course contains 15 students. Figure 2 represents hours of the day for the same course, but a different week. These “hits” are just students entering in the course, not whether they completed any specific task. That type of analytic is in the next section called: Individual Assignment/Content.

Hour of the Day	Hits	Percentage
Sun	12	21.43%
Mon	23	41.07%
Tues	0	0.00%
Wed	10	17.86%
Thur	0	0.00%
Fri	7	12.50%
Sat	4	7.14%
Total	56	



Hours of the Day	Hits	Percent
1	0	0.00%
2	3	6.12%
3	0	0.00%
4	0	0.00%
5	0	0.00%
6	4	8.16%
7	0	0.00%
8	6	12.24%
9	0	0.00%
10	1	2.04%
11	25	44.92%
12	6	12.24%
13	4	8.16%
14	0	0.00%
15	0	0.00%
16	0	0.00%
17	0	0.00%
18	0	0.00%
19	0	0.00%
20	3	6.12%
21	0	0.00%
22	0	0.00%
23	0	0.00%
24	0	0.00%
Total:	56	



Let us explore the figures. Figure 1 (Days of the Week) shows that students entered into the class 56 times over the week, which is about three times for each student. We notice a pattern right away. There is no activity in the third day (Tuesday) and fifth day (Thursday). However, there is more activity in the beginning of the week, then the middle, and reducing near the end.

Figure 2 (Hours of the Day) shows that the peak time of “course access/activity” seems to be around noon, and then peaking in the late hours. Online learning, of course, gives the freedom to work on the course at any time, so we can often see people working late hours.

The goal is not to judge numbers but to understand the numbers and see this class’s activity during this week, and finally, how we can make real-time changes. These changes are based on the course goals.

When looking at these numbers, an instructor must first consider the demographics of their students. One cannot ignore the variety of student populations when comparing online to on-campus learning. Are they working students who have less time to access the course during the week and must access it during the weekends? Are they are more traditional students who can access material during the day (assuming they do not have a full-time jobs)? Are they international, national students or local students?

Some may choose to view this as a microcosm of student activity in a course. This may very well be true, especially in these two examples since the analytics were run during the eighth week of the course. However, remember we are looking at practical, proactive and real-time use of analytics, not as an overall analysis of a course or program.

Based on the results of Figure 1 and Figure 2, how do instructors respond in a practical and proactive method? The use of the data depends on the instructor’s goals in the management of the course:

- a) Is the instructor’s goal to accommodate student schedules?
  - If an instructor understands the demographics of a class and realizes many of them are working students, perhaps they may wish to accommodate the students. Thus the need to reach the class at peak activity arises. An instructor is less concerned with enforcing engagement policies or attendance policy, but more concerned with making it advantageous to students. The reaction to this information gathered should be to make content available at the peak of activity or right beforehand, so the majority of the students view it at once.
  - 
  - Consequently, in the case of Figure 1, it would make sense to post materials at the beginning of week (Saturdays and Sundays). We can also infer that it would be advantageous to post materials, based on Figure 2, during the hours of 11:00am and 1:00pm.

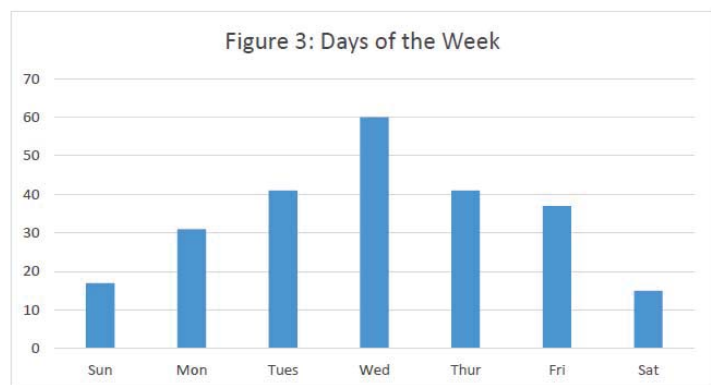
Note: All analytics are in Eastern Standard Time. Adjusting to meet the needs of students in different time zone can be much more difficult.

- b) Is the instructor's goal to create an equal distribution of activity throughout the week, thus promoting engagement to be more dispersed?
- An instructor may be less concerned with students "free time" to do work and more focused on making sure students are following the planned framework set forward in the course. Perhaps a class has a very structured or tiered approach in learning methods. The reaction to the course analytics would be to make the content available at the lowest point of activity. This influences students to attend during the off peak hours.
  - Consequently, instructors may wish to influence the numbers of Figure 1 and Figure 2 by introducing an Interaction Policy.

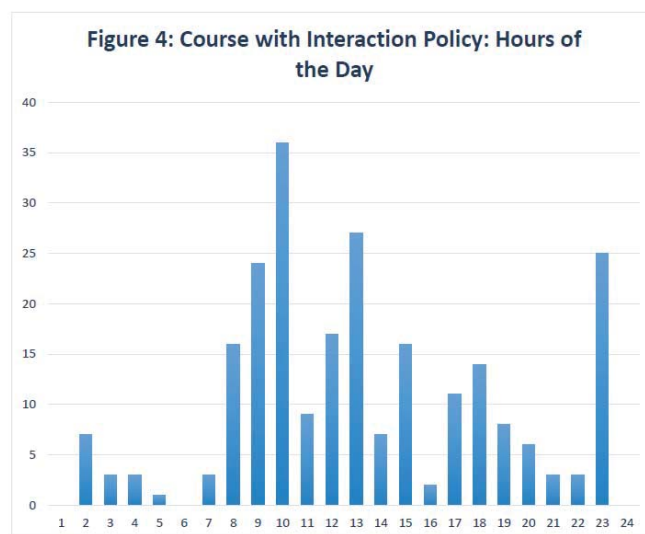
#### 4.1.1 Interaction Policy

Regardless of the management of a course, a common tool in online learning is the global interaction policy. This rule is often part of the graded portion of a course that functions as both an attendance and participation rule combined. It is often a paragraph providing details about how and when interaction will (and should) occur weekly in an online course. An interaction policy may include, but is not limited to: weekly discussion boards, webinars, chat room sessions, video phone sessions, wikis/blogs or any other peer-to-peer learning tools. The instructor can use this policy to reinforce when he/she wishes the student to use the learning tools. Let us look at another course, which utilizes an Interaction Policy. This fully online course contains 25 students and relies heavily on engagement and peer-to-peer learning.

Hour of the Day	Hits	Percentage
Sun	17	7.02%
Mon	31	12.81%
Tues	41	16.94%
Wed	60	24.79%
Thur	41	16.94%
Fri	37	15.30%
Sat	15	6.20%
Total	242	



Hours of the Day	Hits	Percent
1	0	0.00%
2	7	2.90%
3	3	1.24%
4	3	1.24%
5	1	0.41%
6	0	0.00%
7	3	1.24%
8	16	6.64%
9	24	9.96%
10	36	14.94%
11	9	3.73%
12	17	7.05%
13	27	11.28%
14	7	2.90%
15	16	6.61%
16	2	0.83%
17	11	4.56%
18	14	5.81%
19	8	3.32%
20	6	2.49%
21	3	1.24%
22	3	1.24%
23	25	10.37%
24	0	0.00%
Total	241	



This course has a policy that requires students to enter the course at least once before Wednesday and a second time before Sunday. In Figure 4, we see the times of the day activity and also provide a window of hourly (time of the day) activity that reflects certain increased hourly activity. However, in this course, the interaction policy required online students to attend weekly webinars or optional faculty “virtual office hours” that always took place during either early morning or early afternoon (lunch time for many people). This specific course analytic was based on Eastern Time Zone.

What do we notice? Is the Interaction Policy working? What we notice is 242 hits. These hits are a global collection of hits. This means that the students actually just enter the course, not necessarily completing assignments or viewing a piece of content. For now, we see that they are peaking during the fourth day of the week, which is a Wednesday. Remember, the interaction policy required student activity in the course before Wednesday. But we see a lower amount on Saturday, only about 6% of activity. Is this good enough? Is this what the instructor is looking for? The answers to those questions may result in further changes to the Interaction Policy, and continual monitoring of the course analytics. There will probably never be a complete equal distribution of time across the week, but we can begin to gauge how and when students are going into a course.

In this specific case, the analytics seem to show that students are becoming more active in the course weekly. The goal of the instructor, when setting up the interaction policy, was to get students to complete work in discussion boards, open webinar sessions throughout the week and not compress their efforts only into the weekend. However, as with the previous Figure 1 and Figure 2, this is the decision of the instructor to get more activity in the course rather than cater to the needs of student schedules. It’s a balancing act that every instructor must decide upon when building course requirements and now when viewing analytics.

Much of this analysis can seem very focused and specific, but there is still a holistic sense that is always ever present. Instructors cannot ignore the unmeasurable holistic view either. As Campbell, DeBlois and Oblinger (2007) state, “Analytics produces a prediction based on the data available, no prediction can take into account all the possible causes of success or lack of success (problems at home, financial difficulty, and so on).” All instructors address this concern of the level of unpredictability of student populations, so some aspects of analytics will always be holistic. “In addition, some will be skeptical of the ability of an ‘a number’ to account for interpersonal relationships and personal growth that come from attending college or university, irrespective of grades or graduation,” (Campbell, DeBlois and Oblinger, 2007).

## **4.2 Individual Assignment/Content Hits**

The next type of common analytic is called Individual Assignments/Content Hits. This tracks how often students view (hit) a specific content item, such as a document, video, podcast, webinar archive, etc. This analytic can help professors avoid the “wait and see” approach. When a student is studying a concept, there is danger of waiting to see how he/she performs during a quiz. However, this is a “reactive” and not “proactive” approach. The review of this analytic allows for early intervention long before the typical quiz and/or assignment.

Using Pocket Data Analytics, faculty are key to course interventions such as inviting students to office hours, providing additional practice quizzes or encouraging participation in tutorial programs. For some faculty, as Campbell, DeBlois and Oblinger, (2007) state, “analytics may provide a valuable insight into which students are struggling or which instructional approaches are making the greatest impact.”

Let us assume there are bi-monthly quizzes or a monthly quiz preceding weekly lectures. Traditionally, an instructor has only two ways of judging whether students understand the content: 1. Give a weekly or bi-monthly quiz. 2. Ask students: “Do you understand?”

The flaw with these two methods is simple:

- Option 1: the student may perform poorly and their grade is affected.
- Option 2: the student may not necessarily speak the truth or student is not comfortable approaching the instructor for some reason or another (sometimes cultural reasons as well).



The use of the analytic called Individual Assignment/Content Hits allows another approach to supplement the aforementioned options.

Let us look at some more examples. The next two figures are taken from a smaller fully online course but during two separate weeks. Each analytic is also based on two different 40-minute lecture video-casts.

Students	Content Hits
Student 1	13
Student 2	7
Student 3	8
Student 4	6
Student 5	3
Student 6	10
Student 7	3
Student 8	10
Student 9	12
Student 10	9

**Figure 5:** Single Content Item #1

Students	Content Hits
Student 1	10
Student 2	1
Student 3	1
Student 4	1
Student 5	1
Student 6	1
Student 7	2
Student 8	2
Student 9	1
Student 10	3

**Figure 6:** Single Content Item #2

Figure 5 shows how many times each student clicks (hits) this video-cast. What we see is a lot of hits for this specific item from all the students. Figure 6 shows the same group, but this item is only hit once for most of the students. However, one student is viewing this content item 10 times.

There is an understanding that some of the hits may be technical problems or students only watching part of a video. However, that margin of error reduces when the hits reach much higher than 3-4. Regardless of the source of the problem, at first, this should be cause for alarm for any instructor. More than half the students hit the item over four times. Putting aside possible technical problems, what can we infer from these numbers?

1. The content is so interesting that the students are dying to listen to it again and again. Answer: Maybe, but probably not 13 times.
2. Students are struggling with the concepts. Answer: More likely the reason.

Continuing with the use of online course analytics to improve engagement and performance, it makes sense for an instructor to proactively move on this number. Some suggestions may include:

- Create a new topic specific discussion board Q&A
- Create a review sheet for that topic
- Host a topic review webinar
- Host an additional “virtual office hours,” via a webinar or virtual classroom tool
- Create a non-graded review quiz (or even a graded pop-quiz)
- Privately reach out to the student(s) with excessive hits for the content and offer assistance (if needed)

- Do not move onto the next topic until the instructor is convinced his/her students understand the topic.

Regardless of the content type, the instructor must always remember intervention by instructors is an important part of successful online courses. Just because the format has changed, doesn't mean the need for intervention has become any less. As Picciano (2012) states, "In online courses, CMSs routinely provide course monitoring statistics and rudimentary early warning systems that allow instructors to follow up with students who are not responding on blogs or discussion boards, not accessing reading materials, or not promptly taking quizzes. These course statistics are generated in real-time and instructors can review them as often as they wish." The tools are there, but the instructors need to embrace them even on the specific content level.

### **4.3 Discussion Boards**

The third common analytic in an online course is the Discussion Board Forum. Before talking about this analytic, let us talk a little about attendance. Are instructors using this tool as a pure attendance measurement or is there a need to generate a fluid conversation between students in the course?

First, about attendance: absence is a vague and undefined term in online courses. As Dringus (2012) states, "Absence in the online course does not necessarily equate to inactivity, non-performance or no progress by the student. In the online discussion forum, for example, an online learner is a reader as well as a contributor."

Discussion board forums are the most commonly used interactive tool and have been around since online learning started. Its purpose is usually to simulate an in-class discussion, but in an asynchronous method. But how do instructors impress upon students to stimulate an online discussion without forcing them to participate? Perhaps instructors do want to force discussion. Much like other analytics, the management of the course will often aid in the understanding and implementation of changes based on course analytics.

The idea of looking to change habits of students is not a new concept. "Building on decades of research showing that early and frequent assessment is not only a best practice but also a method for changing the studying habits of underperforming students in introductory courses, the team develop an early academic alert system," (Baepler, P., and Murdoch, 2010).

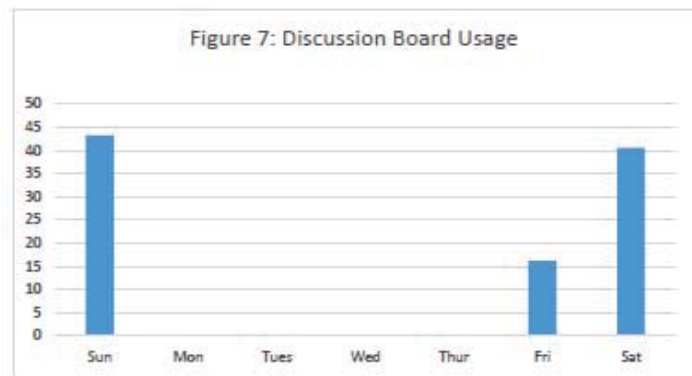
Let us take a look at two examples of discussion board forums from a large fully online course containing 50 students. This specific class is a merged two section course of 25 each. The figures contain a week's worth of discussion board activity.

With figure 7, the discussion seems to be happening on Friday, Saturday, and Sunday, with 0% Monday through Thursday. While discussion board activity is purely instructor's choice, more instructors will probably agree that a discussion board's purpose is to create a fluid conversation over the week(s). It is difficult to have a discussion over three days. However, this also depends on the nature of the course. Assume the discussion is more than simple responses, but is rather part of the course participation grade. The need for a more detailed and lengthy discussion is more prevalent. This specific course does not have a discussion board policy of any kind. The professor was actually unhappy with this discussion board and wished to seek advice for spreading the discussion over the week.

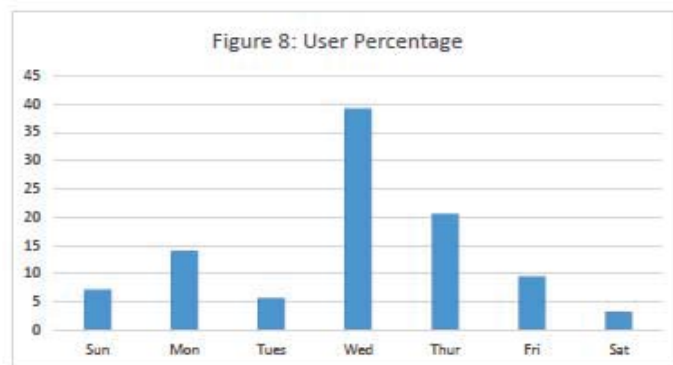
Some of the recommended solutions to improve discussion board activity were to:

- Post follow-up questions (randomly) when the replies are the lowest. The students may be more likely to answer follow-up questions from the professor at random times during the week.
- Create a Discussion Board Interaction Policy, requiring students to post before Wednesday and then another time before another day. This is similar to the global Interaction Policy mentioned earlier.
- Connect discussion board to a major part of the course topics, and thus to a significant part of a student grade.
- Make the discussion board topics extend over more than one week to allow for additional comments.

Days of the Week	User Percentage
Sun	43.24
Mon	0
Tues	0
Wed	0
Thur	0
Fri	16.22
Sat	40.54



Days of the Week	User Percentage
Sun	7.29
Mon	14.1
Tues	5.8
Wed	39.27
Thur	20.6
Fri	9.56
Sat	3.34



The instructor chose the Discussion Board Interaction Policy as part of a weekly grade. Figure 8 was the result during the following week.

As we can see, Figure 8 shows a sudden change in discussion board activity. Now students responded on almost every day (still peaking at the due day for initial posts) but continuing to post the remaining days of the week. In follow-up surveys, the students expressed excitement that the discussion board had become a much more fluid discussion. The professor enjoyed the ability to pull up student activity in a graphical sense rather than swimming through student posts trying to search for a pattern. The running of this simple analytic provided a graphical representation of a discussion that was taking place during a large online course.

## 5. Conclusion

“The trick is to spot low or falling performance early enough to do something about it. Schools often lack the resources to follow an individual student’s progress across subjects from year-to-year. Strong measurement of student performance year after year, combined with predictive analytics, can help highlight factors indicating a downward turn in performance,” (IBM Software Group, 2001). As more and more data is collected from online courses, it only makes sense that instructors use this data to proactively improve engagement and performance. Without the fear of “Information Overload”, the use of this data must be approached with a positive, proactive and ultimately reinforcing mentality. Using the “Pocket Data Analytics” can be a simple process to introduce instructors who may otherwise be hesitant to view data as a predictive part of an online course. Finally, here are five steps for any instructor looking to get started:

- 1) Determine goals behind your use of course analytics;
- 2) Explore your institution’s learning management system to familiarize yourself with the analytics system;

- 3) Begin with a small sampling of course analytics, such as a single assignment or one week of discussion boards;
- 4) Use the methods mentioned in this paper to promote your course goals;
- 5) Once comfortable, expand to other types of analytics; and,
- 6) Always continue to review your goals behind using analytics.

## **References**

- Baepler, P., and Murdoch, C. Academic Analytics and Data Mining in Higher Education. *International Journal for the Scholarship of Teaching and Learning* 4 (2) (2010).
- Campbell, J.P., DeBlois, P.B., and Oblinger, D.G. Academic Analytics: A New Tool for a New Era. *EDUCAUSE Review* 42 (4): 40-57 (July/August 2007).
- Dringus, L. Learning Analytics Considered Harmful. *Journal of Asynchronous Learning Networks*, Volume 16: Issue 3, 87-99.
- Elisa, T. Learning Analytics: Definitions, Processes, and Potential (January 2011).  
<http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf>.
- IBM Software Group. Analytics for Achievement. Ottawa, Ontario, 2001.  
<http://public.dhe.ibm.com/common/ssi/ecm/en/ytw03149caen/YTW03149CAEN.PDF>.
- Olmos, M. and Corrin, L. Learning Analytics: A Case Study of the Process of Design of Visualizations. *Journal of Asynchronous Learning Networks*, Volume 16: Issue 3, 39-49.
- Picciano, Anthony. The Evolution of Big Data and Learning Analytics in American Higher Education. *Journal of Asynchronous Learning Networks*, Volume 16: Issue 3, 9-20.
- Swan, K Introduction to the Special Issue on Learning Analytics. *Journal of Asynchronous Learning Networks*, Volume 16: Issue 3, 5-7.
- Verber, K., Manouselis, N., Drachsler, H., & Duval, E. (2012) Dataset-Driven Research to Support Learning and Knowledge Analytics. *Educational Technology & Society*, 15 (3), 133-148.
- Vernon, Smith. Predictive Modeling to Forecast Student Outcomes and Drive Effective Interventions in Online Community College Courses. *Journal of Asynchronous Learning Networks*, Volume 16: Issue 3, 51-61.

# **PhD Research Papers**



# Exploring Analytics in Health Information Delivery to Acute Health Care in Australia

Chandana Unnithan<sup>1,2</sup>, Katerina Andronis<sup>3</sup>

<sup>1</sup>School of Information and Business Analytics, Faculty of Business and Law, Deakin University, Melbourne Australia

<sup>2</sup>College of Business, Victoria University, Melbourne, Australia

<sup>3</sup>Life Sciences and Health Care, Deloitte Consulting, Melbourne, Australia

[chandana.unnithan@deakin.edu.au](mailto:chandana.unnithan@deakin.edu.au)

[katerina@andronis.me](mailto:katerina@andronis.me)

**Abstract:** The acute health care sector is a “data rich and information poor” environment in Australia. Conversely, information is a crucial, yet underutilized asset for managing patients in health care organizations. To ensure that information being mined and analyzed is of quality and to leverage the power of data analytics tools, a data governance framework need to be in place. The “data concierge” function will provide such a framework for an organisation in analyzing their data. Analytics and predictive analytics are used to manage current and future requirements both from a management and “changing models of care” perspective. Analytics can work hand-in-hand with an organization’s strategic plan that can provide evidence based data and information to support their plan. In this exploratory research investigation, we present scenarios i.e. clinical and non-clinical case studies that demonstrate the use of multiple tools and methodologies for delivery of quality information to acute health sector, in Australia. These scenarios build the case for health data governance. Subsequently, we aim to enhance the understanding on how data governance and analytics can help address current issues in the Australian context of health care, to achieve better information outcomes. In addition to contributing to the body of knowledge, the findings will enable a better appreciation of the analytics and data governance framework, and how it applies to health care practices. Finally, some practical recommendations are offered for establishing and operating analytics and data governance frameworks as well as approaches for justifying the investment for health practices.

**Keywords:** Acute Care, Data governance, Predictive analytics, Data quality, Australia

---

## 1. Background

The Acute Care sector in Australia is experiencing continuous change in Australia in the past few years with implementation of quality systems to achieve accreditation, while maintaining high standards of patient care. The current environment is also characterized by cost restraint balanced with increased accountability. To assist hospitals in providing acute care while also fulfilling regulatory requirements, there is a need for integrated systems and informed solutions (HCI 2014). Conversely, the acute health sector is considered data rich, yet information poor (Duckett & Wilcox 2011). Australian hospitals use ICD10 based codification of data so as to comply with reporting requirements. This data is of such high quality that it can be used not only for understanding current status of patients and for predicting the health in communities, but also help establish appropriate models of care based on predictive analytics techniques.

Specifically, this captured data has the potential to become rich information that can subsequently inform policy frameworks. For example, in a particular area in a region, where there are more reported cases of breast cancer, there may be a requirement for acute care facilities in this area of specialisation. Similarly, if population health data typically on registered births and deaths from cancer registries are synthesised to build a rich picture, this information can predict the incidence of cancer and the continuum of care requirements into the future. Data is a crucial, yet underutilised asset in Australian health care sector, because the rich data obtained from health organisations for managing patients, do not effectively convert into quality information. We hold the view that only when the efficient conversion takes place, and the data becomes high quality information that can feed into policy framework, can it be leveraged for research and care in the acute sector (HCI 2014).

We began this research on the premise that the data that is available currently is not converting into good quality and reliable for predictive analytics to work with. For enabling data conversion into good quality information and reliable for predictive analytics to work with, the function of data concierge comes into the picture. This function enables building good data structures and governance around the data; and then while



predictive analytics can facilitate the evidence based models of care, visual analytics applications such as IBM Cognos can enable better visualisation of the models to inform policy frameworks.

The current aspiration in Australia is to keep people out of hospitals, fit and well within communities. Health facilities on the other hand aim to provide better patient experiences or patient centric care. Leveraging the power of mobile applications, government along with health sector, is trying to facilitate this aim (ACHR 2011). For this purpose, there is a requirement for evidence based models of care and policy frameworks for a continuum of care i.e. tracing a patient journey from a General practitioner, to a specialist, to a hospital, aftercare facilities and home based care. This is where analytics comes into the milieu.

Our exploratory research presents evidence-based scenarios that initially mapped the justification for the data governance. And thereafter, we looked at visualising some of the hospital and health organisation based data governance models that can enable conversion of data into quality information, which can then inform acute health care sector. The rest of this paper is organised as follows. In the next section, we have provided contextual taxonomies. Subsequent section briefs on the methodology and is followed by an analytical discussion of findings. We conclude with some guidelines for the health sector.

## **2. Contextual Taxonomies**

Health Information management is a widely discussed topic although vigorous management of a health organisations data assets is perhaps less pervasive than expected (Andronis & Moysey 2013). There is increased recognition that data is significant to business process outcomes. However, many organizations in the Australian health sector are unsure when it comes to managing health information (or health data governance). Many of the initiatives that do occur are event-driven or short-lived. For example, a sudden change in climate could cause asthma and this may trigger initiatives on chronic disease management. However, post that event, when it is contained, the quality of information collected ceases to be of any interest. Subsequently, unless another such event occurs, the data collected eventually gets archived and abandoned. There is no mechanism in which this data collected becomes quality information that can perhaps prevent the next incidence of this event.

The health sector in Australia is complex with large primary health care providers. The entire clinical overlay with its obvious importance and potential for impact tends to overshadow other information governance perspectives (Andronis & Moysey 2013). This aspect may support the creation and use of clinical data for clinical purposes but generally does not address non-clinical information domains or properly manage clinical information to support broader use. Health care funding, delivery and management are changing in ways that rely ever more critically on management information (Novak & Judah, 2011). Detailed asset utilisation management and strong financial analysis are necessary to properly understand service costs and optimise revenue – critical for financial viability under an Activity Based Funding regime (Novak & Judah 2011). A complete unified set of good quality episodic data is an essential ingredient. However, data is primarily created in the context of individual service delivery processes and is therefore often fragmented and not sufficiently well formed to support the required analysis. Data is a lateral asset spanning multiple functional areas and it is used of multiple purposes in a health organisation and in methods that may not be significant in the context of where it is created. The effective management is challenging, as the lateral characteristic does not align well with management requirements.

An insight into models of health care is presented as a prelude to understand the context better. Models of Care within Australia define the way health services are delivered. These outline best practice care and services for individuals or population groups as they progress through the stages of a condition, injury or event. It aims to ensure people get the right care, at the right time, by the right team and in the right place (Duckett & Wilcox 2011). Essentially, the models describe typical activities to be delivered to patients by a provider, health professional or care team; type of services to be provided by an organisation; the appropriate stage for an activity or service to be delivered; the location or context that the activity or service will be provided in; the health care team and community partners that will provide the service; and the policy framework for the model of care (ACHR 2011). These models are based on clinical specialties and are developed by clinical teams collaborating in providing care pathways for the management of chronic diseases, trauma cases etc. Every Health Department in Australia provide state-wide framework of “models of care” for

different conditions. Health facilities adopt these and modify them as required to “fit for purpose” for the care of their patients (Duckett & Wilcox 2011).

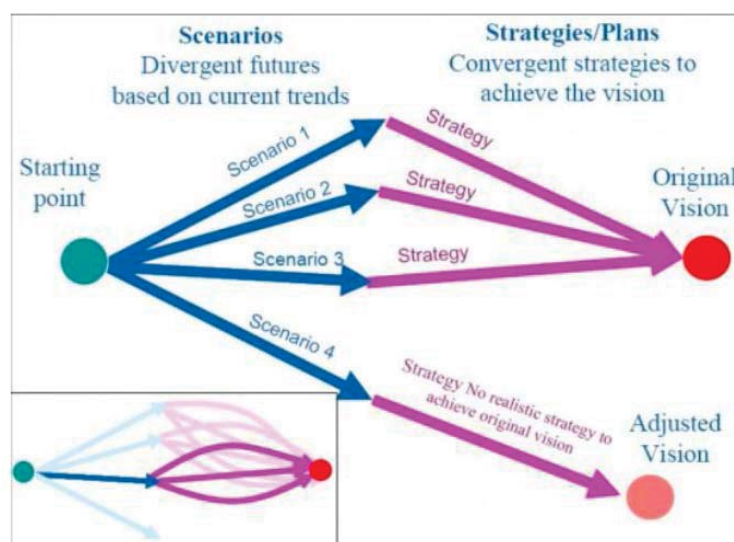
The Data Management Association of Australia (DAMA 2014) defines data governance as the exercise of authority and control (planning, monitoring and enforcement) over the management of data assets. The International Association of Information and Data Quality (IAIDQ 2014) purport that data governance is the management and control of data as an enterprise asset. Drawing from these major definitions, we have applied the framework provided by IAIDQ (2014) to analyse and visualise the health care scenarios in section 4.2.

Within the data governance framework is a data concierge function that manages the information highway for data stakeholders, owners and users (IAIDQ 2014). This is a very important function as it manages the information that comes from different environments that need to be analysed and standardized before it enters the enterprise hub. Requirement teams (data concierge’s) work together to standardize data to ensure information moving forward which is based on standard quality. This is imperative when using data for clinical research and clinical decision making that everyone understands that the data elements and they are all on the “same page”. To achieve this a data concierge team provides the following functions:

- Data custodians work with the data stewards and data owners to incorporate data.
- Data stewards work with the business users and data owners to standardize the data
- Business users work with data stewards and data owners to define business requirements
- Data custodians and application services work with architects and business users to develop solutions
- Data owners and data stewards work with the business users to maintain the data.

### 3. Methodological Framework

In this research investigation, we have used the concept of Scenario building and analysis towards building a strategy (Creswell 2013). Throughout the next section, we provide scenarios and plans adjusting the vision as required gradually.



**Figure 1:** The process of Scenario Building, Analysis and Planning (adapted from Moriarty et al, 2007).

In a similar context, Bierbooms et al. (2011) performed a scenario analysis of the future residential requirements for people with mental health problems in Eindhoven. There were four steps involved: 1) an exploration of the external environment; 2) the identification of key uncertainties; 3) the development of scenarios; 4) the translation of scenarios into guidelines for planning organizational strategy. To explore the external environment a document study was performed, and 15 semi-structured interviews were conducted. During a workshop, a panel of experts identified two key uncertainties in the external environment, and formulated four scenarios.

Taking this study as a guideline, this research was performed with following steps: (1) we explored the environment; (2) identified the key issues in data governance; (3) developed scenarios based on them; and, (4) translated the scenarios into the contextualized guidelines and framework for data governance. In exploring the environment, 20 semi-structured interviews were conducted over a time period of 3 years. A panel of experts then identified the key issues in a workshop and formulated the six scenarios, which was further validated individually with a set of one-to-one interviews with experts by the researchers.

## 4. Findings and Discussion

In this section, we present our findings and an analytical discussion based on a scenario analyses, funnelling into conceptual models that represent data governance interaction with health practices in Australia.

### 4.1 Scenario Analysis - The need for Data Governance

The series of information related scenarios are drawn from the primary health care provider contexts and these cover clinical and non-clinical areas in order to demonstrate that there are significant dependencies and impacts for data across the board. Each scenario is considered from the perspective of how effective data governance contributes to a better outcome. The purpose is to show how data governance generates improvement. Although the discussion here focuses on the data governance related aspects, in all cases there are other factors in play as well such as related process improvement for acute care management.

#### 4.1.1 Scenario 1 – Scrutinising the business case of an operating theatre in the State of Victoria

An analysis was conducted of theatre demand and usage including forecasting. The conclusion was that investment in another theatre was warranted with an expected payback period of five years. After construction of the theatre was well under way, the analysis was revisited and found to be incorrect due to misinterpretation of the data. Corrected analysis showed that the return on investment for the additional theatre would take five years longer to achieve than expected.

This scenario is one of conducting analysis that underlies an investment business case. This would require data representing past and likely future operational volumes and demand, case-mix and cost and revenue elements. Such data is likely to come from a range of sources and needs to be correctly understood and applied in order to get a valid analysis outcome. A sound data governance regime would facilitate identification of the people who really understand the required data, where to get the data from and the associated definitions. It is also likely to improve the chances of consultation with the appropriate people to understand the business aspects relevant to such an analysis and ensure that the analysis is properly validated before being acted upon.

#### 4.1.2 Scenario 2 – Information focus in a public primary care provider

In public primary care in Australia the information focus is on in-patients where there is generally comprehensive record keeping. For outpatients (ambulatory) however, other than basic episodic data, record keeping is relatively poor as it lacks a longitudinal view and health service outcome details. Although coding outpatient data would involve significant costs, without such detailed data a provider cannot get a precise picture of their service and cost base nor understand their *casemix* for management or clinical research purposes.

Large public hospitals (and care providers generally) are subject to significant data collection obligations that include critical submissions that directly affect funding. Complying with these collections generates a substantial workload. Most (perhaps 80%) of the data required for these collections is drawn from data needed to deliver clinical care and support administration. Effective data governance contributes directly to ensuring data is correctly captured, sourced and reported; and that creators of data (including clinicians) understand what data is required and why. Executive sponsorship would ensure that this activity has an appropriate focus given its importance to the organization. One could also expect better overall efficiency through taking a more holistic view of data and collection and reporting processes. The quality and completeness of the reporting would be improved reducing the rework associated with submitted data that fails validation checks.

#### *4.1.3 Scenario 3 – Influence of genomic data*

There is increasing use of genetic data for characterisation of pathogens and tumours, and their application for biomarkers in patients, that directly influences clinical decisions on treatment. The genetic data itself and translational correlation between genetic markers and treatment outcomes has intrinsic value. Relevant data sources are often spread across multiple systems that are generally not linked, but linking the data is the key to unlocking its clinical value. Health outcomes for the community will be improved by sharing these data. This creates an information management context that is very different to the past.

The use of genetic information is significantly increasing the quantity of data to be managed by care providers. It also requires sophisticated indexing and linking with other clinical and demographic data in order to be used effectively. There is a great opportunity to improve clinical outcomes but only with corresponding data governance. There is little point (and perhaps risk) in influencing clinical decisions with incorrect or incomplete data or data that has been incorrectly linked. Exact compliance is needed since the outcomes can be specific to the patient and the patient's circumstances. Precision is essential and is only achievable with appropriate data governance to help manage the quality of the underlying data, its interpretation and apt use.

#### *4.1.4 Scenario 4 – Optimizing Income*

With the shift to Activity Based Funding, care providers now find themselves being paid on the basis of “delivered units of care”. This necessitates a detailed understanding of care delivery costs down to episode level, careful selection of treatment and payment options for patients, and diligent clinical coding to maximize funding outcomes. Much of this lies outside the purely clinical information domain yet is critical to delivering the most effective health outcomes and simply staying in business. The collection of accurate data involves the whole patient episode journey and relies upon clinicians entering complete clinical data (including complications) into the patient's notes. This facilitates the accurate coding required to achieve optimised revenue outcomes for the hospital. There is often a material level of data governance around this coding process that is largely focused on the clinical rather than administrative significance.

The dramatic shifts in funding approaches to output and volume based models requires care providers to have an accurate and detailed handle on costs and be able to view costs from perspectives that include clinical characterization, outcome, patient, service and time. Data governance needs to extend beyond the clinical domain and support more than clinical costing. Ensuring that revenue is diligently managed has become a critical capability. Given the reliance this has on data codification, integration and interpretation, data governance is an essential enabler. It also enables essential knowledge to be captured and harnessed for re-use rather than locked away in the heads of individuals.

#### *4.1.5 Scenario 5 – Campus, Network and Regional*

In Australia, health care services have undergone dramatic change from individual funded and managed sites to networked groups of sites funded collectively. In addition, there is now a regional public health primary care overlay that is responsible for optimising the health care services in each of 60+ areas across the country. While these management structures are intended to improve efficiencies and better align services with demand, there is an underlying requirement to be able to report information consistently across these organisational structures as well as drill down to a detailed level by provider, site, service, patient type etc. The information structures and facilities required to support this is not sufficiently mature.

Management reporting requirements for a large health care provider can be complex, in keeping with the nature of the organization and the complex environment in which it operates. Structural changes that suddenly deal with groups of previously separate organisations as unified networks of campus based organisations have generated significant challenges in management reporting. Data governance is necessary to define terminology and agree business rules that enable common reporting frameworks and consolidation of reporting elements across multiple organisations. This is essential to meet external commitments. But it is also essential internally (within a network) simply to make sure that data available to management is actually correct and provides management insight into performance at a granular level. This is especially important since a major premise for the structural changes is that they yield efficiencies, not simply through scale or aggregation but by forcing optimisation. Making really important management decisions about optimisation

(such as service portfolio changes) requires reliable information based on reliable data across sources of data that have previously not been integrated.

#### 4.1.6 Scenario 6 – Cancer research information exchange framework

A cancer centre in Australia is currently embarking on a journey to implement a clinical and research information exchange framework. This framework would enable clinical information sharing across multiple hospitals for clinicians to make informed decisions for their patients as they move from one health facility to another. This project aims at implementing a research information exchange so as to share their research data with local and international research partners. To enable this, currently a data governance framework has been established with a data concierge function that allows information to be shared in a secure manner for the development of clinical and research outcomes to assist in the cure of cancer.

The reason for effective data governance in this instance is to create a rapid learning environment for cancer that is founded on trusted data sets, for those that support integrated cancer care across there Cancer Centre partners or those that enable the creation of associations, rapid learning and ultimately high impact cancer research. Delivering trusted data sets requires an effective data governance framework, encompassing the roles, policies, frameworks and tools that collectively deliver high quality and trusted data needed to deliver a successful rapid learning environment for cancer.

The experience of Cancer organisations around the world recognise as having achieved a significant measure of success in the management of the data is actually the single most important success factor. Suitable technology and software applications are necessary to provide capability but can only do what the data allows. The true business of the Cancer Centre itself is the management of data and therefore data governance constitutes a core management discipline and key operational processes. Given the nature of the data involved, the Cancer Centre also has significant professional and regulatory responsibilities to comply with in relation to acquisition, storage, access and use of this data. The key point though, is that data governance is central to materialising the Cancer Centre’s vision, not just something that is necessary for compliance.

## 4.2 Visualising Data Governance in Health Practices

In the context of Australian primary health care providers, Andronis & Moysey (2013) conducted an investigation based on a structured framework (IAIDQ 2014) to assess the indicative level of maturity of data governance. The framework posed questions derived from the content of 4-6 performance domains and focusing on data governance as illustrated in figure 2.

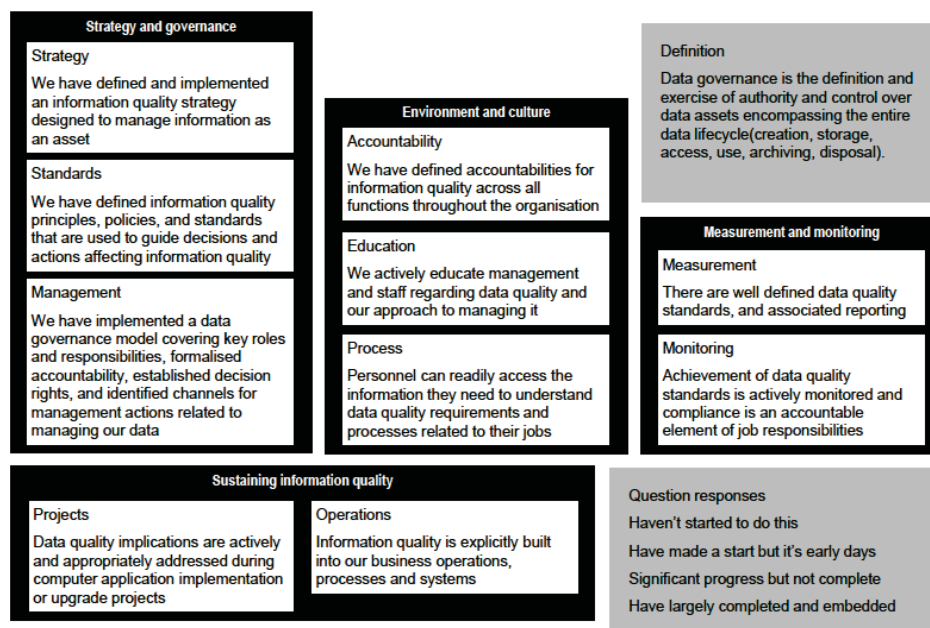
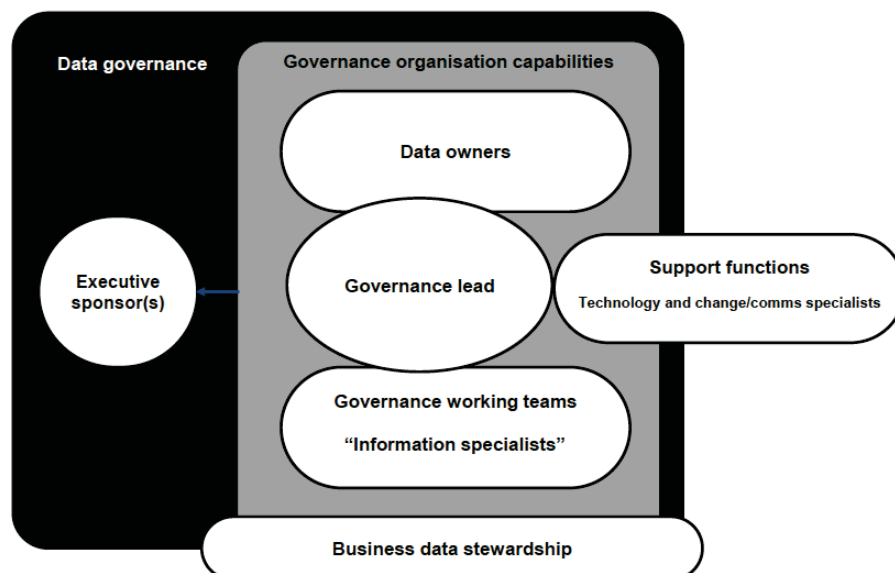


Figure 2: Structured Framework for assessing Data Governance Maturity

The key pattern indicates that overall maturity for data governance was quite low across all performance areas. Where it was relatively mature, it focused on aspects of privacy or clinical records where there was action required. None of these warranted treating data as a crucial asset. While this does not mean that data governance itself is uncontrolled or patient outcomes are at risk. Rather, the effect is a less obvious combination of multiple versions of data, poor quality data, loss of efficiency and relinquished opportunities.

In Australia, there are significant changes in health and hospital administration and funding that have major implications for the use of data in reporting and management. Today, individually managed and operated sites which were funded on size have been grouped into local health networks that operate as single multi-campus entities funded on healthcare outputs; their reliance on data to function thus has magnified. There is also additional government overlay in the form of 60+ local administration areas known as Medicare locals to better integrate primary health services on a localised basis. This generates a requirement for detailed data covering many services that are previously relatively scantily documented. The government administrators on the other hand, want to make evidence based decisions, which increases the need for complete granular data. Many data intensive situations require the ability to combine data from multiple sources and therefore, aligned definitions and compliant data recording and encoding (Andronis & Moysey 2013).

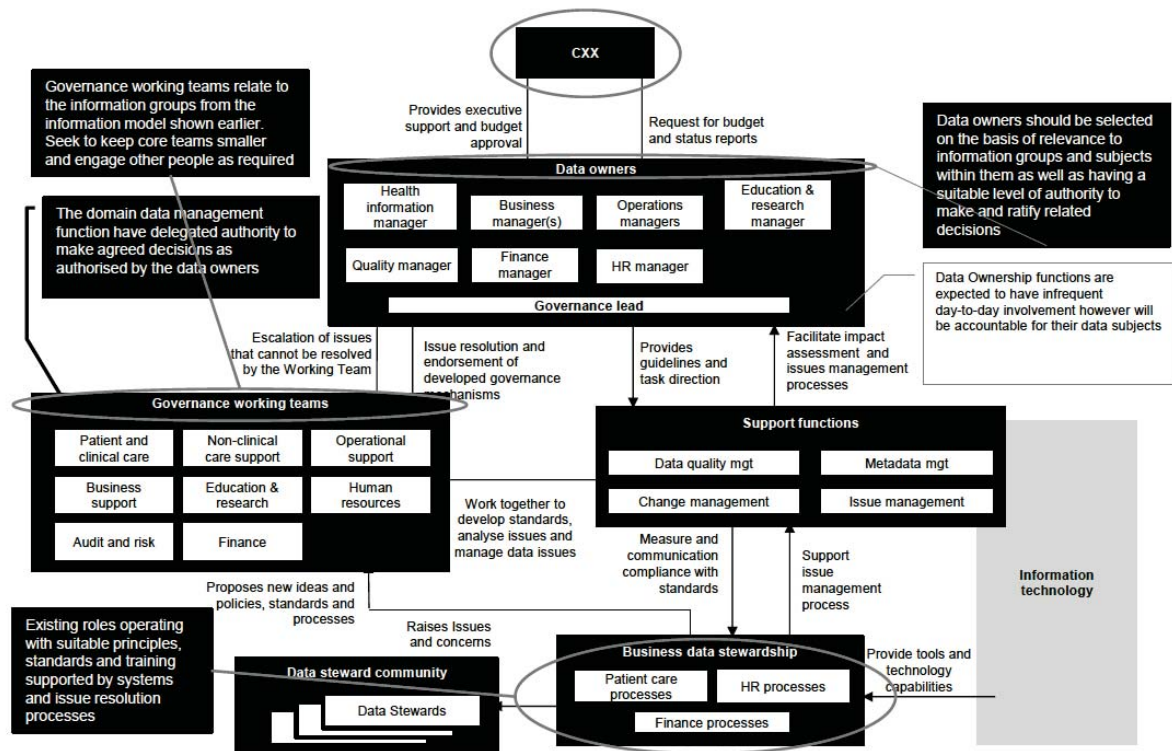
In the health environment, the key interest is to reuse once entered data many times over. However, this is rather challenging in an environment where disintegrated systems co-exist. Andronis & Moysey (2013) presented a high level model for data governance, as presented in figure 3.



**Figure 3:** A high level data governance model (Adapted from Andronis & Moysey 2013)

Subsequently, the authors applied this model to the context of a large health provider as represented in figure 4.





**Figure 4:** Data Governance model applied to a HealthCare provider (Adapted from Andronis and Moysey 2013)

In the figure 4, the application of the proposed data governance model to a healthcare provider is depicted and explained as follows.

**CXX – Executive Sponsor:** The executive sponsor has overall responsibility and authority for (in this case) data governance and exercises this through control of a group of line managers responsible for different business and clinical areas.

**Data Owners:** A Data Owner typically should be someone who is responsible for an area of the business that is a key user of the data domain. However, an important consideration for this role is the management of a data domain on behalf of the organisation, not just for the area of direct responsibility.

**Governance Working Teams:** Governance working teams that align with data domains, recognising that knowledge of a data domain is usually spread across multiple individuals.

**Business Data Stewards:** Explicit inclusion of business data stewardship to recognise that day-to-day responsibility for creation of data rests with the people at the “coal face” of business and clinical processes.

## 5. Recommendations and Outlook

In the health context and specifically acute health care, adhering to some recommendations can improve the data quality and improve health outcomes.

- Data quality has to become a common goal across all performance areas. For this purpose, it is best to adopt a framework, illustrated in this paper, to provide the context of care for each activity undertaken. This can result in convergence of good quality data that can result in better patient outcomes. For example, good quality data can be used for continuum of cancer care and support of research for curing cancer.
- Do not resource manage information governance as a project. Rather build the governance arrangements to regular business of the health practice as such. Ensure team participation, as data governance is a team-based game. It can only be effective when stakeholders understand the purpose, know their roles and feel that they are on board.



- Execute modestly scoped activities that are of clear priority. While planning for data governance is reasonable, it should be recognised that you will only move forward where there is relevance and support.
- Take steps with permanence in mind, as a once-off improvement in data quality achieves almost nothing. The adage that “what gets measured gets managed” is true when it comes to data quality. Even very basic data quality reporting helps instil it as a relevant business activity.
- Keep it simple – for example just go for a basic data dictionary in the first instance. Implement the governance organisation as data subjects are addressed and Implement data quality reporting as subject areas are addressed.

In summary, Australian health care and acute care framework has miles to go before the desired data quality is achieved. We have visualised some scenarios, a framework that can be applied to healthcare practices in this paper. After the e-health records implementation occurs nationwide, Australian healthcare sector needs to strive to achieve data quality, using data governance and concierge techniques reviewed in this paper. Only then, the distant dream of acute health care management and continuum of care for Australians within their homes can become a reality.

## **Acknowledgement**

The lead author Chandana Unnithan, as an academic staff member of Deakin University – School of Information and Business Analytics, Faculty of Business and Law and a PhD candidate at Victoria University – College of Business, wishes to thank both Universities and the staff, for their support with completing this project successfully.

Katerina Andronis is the Director, Life Sciences and Health Care, Deloitte Consulting, Australia. Both authors’ acknowledge Deloitte Consulting for their support with this project.

## **References**

- ACHR (2011) Healthcare in Australia - Prescriptions for Improvement, Australian Centre for Health Research, South Melbourne, Victoria.
- Andronis, K and Moysey, K. (2013) Data Governance for Health Care Providers, a chapter in *Health Information Governance in a Digital Environment*, Studies in Health Technology and Informatics, Volume 193, IOS Press, Australia.
- Bierbooms, J.P.A.J., Bongers, M.B.I., Oers, A.M.H. (2011) A scenario analysis of the future residential requirements for people with mental health problems in Eindhoven, BMC Medical Informatics and Decision Making Journal, 11(1), BiomedCentral.
- Creswell, J.W. (2013) *Research Design: qualitative, quantitative and mixed method approaches*, 2<sup>nd</sup> Ed, Sage Publications, Thousand Oaks, California.
- DAMA (2014) [www.dama.org.au](http://www.dama.org.au) [online]
- Duckett, S.J. and Willcox, S. (2011) *The Australian health care system*, Edition 4, Melbourne: Oxford University Press.
- HCI (2014) Acute Care, Healthcareinformed.com, [http://healthcareinformed.com.au/acute\\_care.html](http://healthcareinformed.com.au/acute_care.html) [online]
- IAIDQ (2014) <http://iaidq.org> [online]
- Moriarty, P., Batchelor, C., Abd-Alhadi, F., Laban, P. and Fahmy, H. (2007) *The EMPOWERS approach to Water Governance: Guidelines, Methods and Tools*, Presented at Inter-Islamic Network on Water Resources Development and Management (INWRDAM), Amman, Jordan.
- Novak, J. and Judah, A. (2011) *Towards health productivity reform agenda for Australia*, South Melbourne: Australian Centre for Health Research.

# Information Visualisation and Knowledge Re-construction of RFID Technology Translation in Australian hospitals

Chandana Unnithan<sup>1,2</sup>, Arthur Tatnall<sup>2</sup> and Stephen Burgess<sup>2</sup>

<sup>1</sup>School of Information and Business Analytics, Faculty of Business and Law, Deakin University, Melbourne, Australia

<sup>2</sup>College of Business, Victoria University, Melbourne, Australia

[Chandana.unnithan@deakin.edu.au](mailto:Chandana.unnithan@deakin.edu.au)

[Arthur.tatnall@vu.edu.au](mailto:Arthur.tatnall@vu.edu.au)

[Stephen.burgess@vu.edu.au](mailto:Stephen.burgess@vu.edu.au)

**Abstract.** Radio Frequency Identification (RFID) is an evolving technology innovation that uses radio waves for data collection and transfer without human involvement. With its success worldwide in hospitals for improving efficiencies and thereby quality of care, the technology was piloted in Australian hospitals in late 2000s. However, existing literature (in 2013) reflects limited success in a full-scale implementation and emerging view that the socio-technical factors in implementation are not being considered. Information systems researchers in Australia had begun emphasizing socio-technical approaches in innovation adoption and translation of technology in the context. A qualitative research study, with a multiple case study method was set in this premise in 2007 and aimed at addressing the knowledge gap. Information for the case studies was obtained through a rigorous data collection through semi-structured interviews, focus groups, concept mapping and documentation analysis. The findings were then validated for currency with practitioners in the field in 2013. Innovation Translation is an approach that regards that any innovation needs to be customised and translate in to context, before it can be adopted. To understand the ‘social’ aspects that may be involved in adoption, an Actor-Network Theory informed lens helps reconstruct the implementation process - investigating social networks and relationships that influenced innovation translation. The Innovation Translation approach to theorisation and knowledge abstraction, informed by Actor-Network Theory (ANT), removed the need for considering ‘the social’ and ‘the technical’ in separate modes. More importantly, an ANT informed lens acted as an augmented filter enabling in-depth view of the data and information visualisation of knowledge abstracted in this research investigation. The contributions of this research are in that it addresses the socio-technical gap evident in academic literature pertaining to RFID technology adoption in Australian hospitals; augments the body of knowledge concerning innovation translation in health informatics informed by the Actor-Network theory analysis. The ANT lens visualises that RFID was facilitating the re-negotiation and improvement of network relationships between the people involved, as well as the technology. The knowledge abstraction and visualisation enabled by ANT is a major contribution to the field. From a practical perspective, it is evident that RFID innovation translation requires commitment of a key stakeholder in Australian hospital operations who is able to influence others, a view that is validated by industry practitioners.

**Keywords.** Information Visualisation, Knowledge abstraction, Actor-Network Theory, RFID, Innovation Translation, Socio-Technical Approaches, Hospitals, Australia

---

## 1. Introduction and Research Motivations

Technology adoption in hospitals is always challenging as also reflected in the reviews from the last few decades (see for example Coustasse, Tomblin & Slack 2013; Yao, Chu & Li 2012). Academics have used adoption models, such as the technology acceptance model (TAM), which evaluates user acceptance of computer-based information systems (Davis 1986); or diffusion of innovation (Rogers, 2003) to study and evaluate technology adoption within hospitals.

Hospitals are no doubt, are chaotic environments where regular scheduled processes undergo rapid changes in case of emergencies. For example, in case of an emergency it is not uncommon to page all available nursing staff. All procedures that were scheduled, with the exception of critical surgeries, could be postponed at this time. Clinicians posted in the outpatient area may be summoned to assist with emergencies. In this chaotic environment, technology cannot be implemented using the standard procedures or techniques used in other environments, such as retail or manufacturing. While literature is prolific on the technical, legal and economic impediments of technology implementation, the rather hectic social environment of hospitals (in which technology is implemented, making it unique) is often ignored. In this research, we take the view that technology implementation in hospitals involve a unique challenge. In industry sectors such as retail or manufacturing, a routine adaptation is undertaken to redesign processes, and users are then trained to fit in with the new technologies. However, health services cannot be interfered with, stalled or put on hold temporarily: this may cost a life! Given the perilous nature of hospitals, we were of the view that any new

technology has to be customised and implemented with the involvement of all ground level users or stakeholders.

Radio Frequency Identification (RFID) is regarded an accepted mobile technology solution to improve process efficiency in supply chain management. It uses radio waves for data collection and transfer, efficiently and automatically without human intervention (Yao, Chu & Li 2011). From 2005, health care providers globally began to realise the benefits of adopting RFID into their operations, to enhance efficiency and provide better services (Degaspari 2011). By 2010, RFID systems had been trialled to track medical equipment and supplies more efficiently in hospitals (Yao, Chu & Li, 2012) as RFID was expected to lead reduction in clinical errors, reduced costs and increased efficiencies.

A literature review from 2000–2013 (Coustasse, Tomblin & Slack 2013) ratifies the acceptance of RFID as an innovation accentuated by global recession that called for cost efficiencies, without compromising on quality of care. Nonetheless, in the Australian context, privacy and health industry regulations have mainly constrained the innovation diffusion. Australian hospitals are still transitioning into e-health records and unified health systems (Dunlevy 2013). RFID integration issues with legacy clinical systems and the costs involved in large-scale implementation resulted in hospitals putting RFID on hold as they prioritised funding towards e-health records transition (Duckett & Wilcox 2011, Novak & Judah 2011).

The motivation for this investigation arose in 2007 from RFID's status as a still nascent innovation in Australian hospitals. There were only a small number of largely unsuccessful and abandoned cases or pilot studies reported in Australia involving large hospitals, through biased vendor or technical reports. Even in 2013, focused research reporting on RFID adaptation in Australian hospitals is yet to emerge and limited to typical diffusion studies or uses essentialist approaches, which are indifferent to or dismissive of socio-technical factors (Dunlevy 2013). Therefore, this study was and remains significant to-date because the findings throw light upon the under-explored, yet critical interplay of, socio-technical factors that have a prominent role in the adaptation of RFID technology in Australian hospitals.

Conversely, academic literature in Australia was demanding more qualitative approaches and interpretive studies. In the early 1990s that influential researchers in Australia (for example, Parker, Wafula, Swatman P & Swatman PMC 1994) alluded to the fact that information systems research has to move away from technical issues and focus more on behavioural issues, and IS methodologists called for a move away from scientific/positivist research methods towards interpretivism. As Davis, Gorgone, Cougar, Feinstein and Longnecker (1997) pointed out, the problem of handling complexities due to interconnected technologies, how they relate to humans and organizations, and how humans relate to them, had come into focus. However, as Tatnall (2011) pointed out, even interpretive studies had focused on innovation diffusion and technology related factors, taking on an 'essentialist' approach that suggests that innovation diffusion occurs because of the technology's salient features making it acceptable. These approaches, such as diffusion studies or TAM, did not reveal possible and deeply covert human and social factors and perceptions. In other words, the quintessence of 'innovation translation' (i.e., how RFID as an innovation can translate into Australian hospitals, where 'implicit socio-technical factors' play a significant role) was rather deficient from published academic literature.

Taking the gaps in literature into account this investigation took a holistic approach, wherein we observed and analysed the adaptation process of RFID in two different large hospitals, with several departments: this proved revelatory. The study, which has evolved over five to six years, made significant discoveries through participant and non-participant observations. Further, strong validation from practitioners in the health sector (hospitals, health consulting and professionals), on the findings and recommendations, in 2013, has made it relevant to the current situation. Moreover, the study not only applied the innovation translation concept to the field, but also visualized the interplay of human beings and the technology, using the Actor-Network Theory (ANT) Lens. ANT has been criticized for its insufficiency in explaining the relationship formation between actors and changes that occur in relationship networks (Tatnall, 2011). The investigation addressed this deficiency in relation to information systems research, in that it incorporated an ANT lens for visualisation of the innovation translation theory as an augmented filter, enabling an in-depth view of the data, thus making an important academic contribution to the field of ANT.

The rest of this paper is organised as follows. In the next section, we provide relevant literature and the formation of the conceptual framework using theoretical constructs. Subsequently, we brief on the methodology and research design in the following section. This section is followed by our findings discussion which has been presented using the ANT lens funnelling into conclusions i.e. knowledge abstraction and implications of this research.

## **2. Relevant Excerpts of Literature and Theory Framework**

Towards the end of the 1990s, the application of socio-technical perspectives was arguably promoted as a means to appreciate and extenuate the poor uptake and performance of information systems within health sectors (Berg 1999). Emerging socio-technical foci also seem to have influenced research, as more qualitative studies began to emerge alongside conventionally quantitative-focused global health sector research (Whetton & Georgeou 2010).

### **2.1 The Australian Health Milieu**

Over the past decade (2001–2010), it is evident through the progress of varied organisations, such as Health Informatics Society of Australia, National E-Health Transition Authority (NeHTA) etc, Australia is progressing towards a fully electronic health record system nationally (Prgoment, Georgiou & Westbrook 2009), albeit slowly. Mobile technologies are the fastest growing category of the ICT revolution. Both public and private health care providers are increasing their investment in technology, particularly in mobile communication, to enable process efficiency in their workforces (Ho 2012). Debatably, two of the largest issues facing hospitals are enhancing worker productivity and reducing human error (Ho 2012).

It was only towards end of 2006 that mobile technologies were initially explored in Australian hospitals (Cangialosi, Monaly & Yang 2007). In 2009, with Wi-Fi evolution, wireless devices and mobile technologies like RFID gained momentum in Australia (Yao, Chu & Li 2012). Unlike other technologies, which may address a particular area (such as physicians or nurses with a PDA), or a technological system that is meant for physiotherapy or surgery, RFID is a tracking technology that has the potential to track objects and people in a hospital. By the sheer nature of its tracking ability, it has the potential to pervade hospitals, touching every department dealing with patient care. It was in the decade 2001–2010 that technology refreshments had begun to occur in earnest, with the imminent national health records system. Additionally, in Australia, health care is heavily affected by privacy regulations (Privacy Act of Australia 1988; Privacy.Gov 2013) and the Privacy Act is more formidable. Often it is so doctrinaire that any data regarding an adult patient is not even provided to parents, unless they are named as carers, and if the patients are unable to handle themselves. This poses significant difficulties in emergencies (Duckett & Wilcox 2011). In such restrictive conditions affecting the health sector, a piece of technology, such as RFID with its surveillance potential, was unacceptable in its original form to Australian hospitals. They had to be compliant with the existing privacy laws and standards.

RFID, as an innovation, was still being trialled towards the end of the 2000s, and was not yet fully accepted as a standard way of asset-tracking in hospitals (at the time when this research study was conducted, i.e. 2007–2012) (Duckett & Wilcox 2011). An early report of RFID-enabled functions in Australia by Bacheldor (2006), reported that the Rockhampton Base Hospital in Queensland, Australia, used RFID to improve nurse safety in mental health ward buildings. Many pilot implementations, such as those at RMH and Barwon Health in Victoria (and others in Western Australia and Queensland) were introduced for not only asset-tracking but also patient tracking (infants, geriatrics & mentally disabled). However, most of these pilots did not result in full-scale implementation. Bendigo Hospital in regional Victoria has successfully deployed RFID progressively in all departments across the hospital (Friedlos 2010). Subsequently, they have planned an in-built RFID-enabled infrastructure for the new hospital building to open doors in 2013. Despite Australian hospitals piloting the technology towards the end of the decade, unsuccessful and abandoned examples persisted.

When the socio-technical factors are considered, research is still emerging. RFID is almost a non-interventional technology in the health sector for tracking equipment, so it may be accepted for that purpose (Coustasse et al., 2013). However, the socio-technical factors, or actors in the milieu (stakeholders), hold the key to effective RFID deployment in the Australian context. Within hospitals, a business case might encourage management to consider the technology; however, it is rendered useless if abandoned by users. Therefore, in this research we investigated the socio-technical factors and how the interplay between RFID technology and actors (people) affect successful deployment. The research questions were as follows.

RFID technology in Australian hospitals is still nascent and seemingly viewed with scepticism:

- What socio-technical factors interact to affect the adoption of RFID in Australian hospitals, and how is this achieved?
- What are the key factors that affect adoption?
- How do these factors interact and negotiate to eventuate in adoption?

## **2.2 Developing the Conceptual Framework**

Callon (1986) outlined a novel approach to the study of power, the 'sociology of translation'. This pivotal work describes a scientific and economic controversy on the causes for the declining population of scallops in St. Brieuc Bay, and the efforts by three marine biologists to develop a conservation strategy for that population. We have drawn from this seminal work to build the conceptual framework described as follows. The research quest was to determine the socio-technical factors that affect adoption of RFID in Australian hospitals, and how they interact to eventuate successful translation.

- Problematisation is where a group of key actors define the nature of the problem (in this situation, the issue that RFID proposed to address) and the role of all actors in the context. All factors in this situation are regarded more as 'actors'.
- Interessement is whereby the factors or actors defined in the problematisation process impose the identities and roles defined on other actors, thus building a network of relationships where all actors become involved. In this context, the champions of RFID in the context try to negotiate with others in the network.
- Enrolment occurs after the success of Interessement, when a process of coercion, seduction and consent leads to the establishment of a stable alliance. In this situation, this is the moment of translation whereby the key network of actors enrolls all others into accepting an RFID solution, either by imposing it or through influencing them.
- Mobilisation occurs when the solution gains wider acceptance. In this situation, RFID gains wider acceptance as a solution for the proposed reason, within the hospital context.

Thus, we took an approach to theorising innovation that had the advantage over essentialist approaches, such as innovation diffusion.

## **2.3 Intersection of Innovation Translation and Actor-Network Theory (ANT)**

In the much-acclaimed diffusion of innovations theory (Rogers 1995, 2003), an innovation is defined 'as an idea, practice or object that is perceived as new' (Rogers 1995: 11). Tatnall (2011) purport that the process of innovation involves getting new ideas accepted and new technologies adapted and used. A challenge to adapting a new technology, or applying it to an existing context, is that it is often not accepted in the form first proposed (Latour (1986, 1996) and have to be translated into a suitable form for the environment. Unlike diffusion, 'translation' can only occur if it interests all involved stakeholders.

In this research, it was studied how RFID technology translated into large Australian hospitals, retaining the essential element of innovation, namely 'location tracking', while adapting itself to the environment. Elements of ANT were incorporated into the conceptual framework, thereby visualising the interplay that occurred between actors. ANT was developed by Latour (1986), in an attempt to give a 'voice' to technical artefacts, considering humans and non-humans as equally important in the translation of innovative technologies. As Tatnall (2011) purports, ANT offered an advantage over other theories, as there is no dividing line between human and non-human entities, nor have an essence attributed to either that determines the adoption rate of an innovation. However, one criticism of ANT is its inability to explain relationship formation between actors, and over changes of events in relationship networks (Greenhalgh & Stones 2010). In this research, this criticism has been addressed in relation to information systems research, in that the conceptual framework incorporates an ANT lens for visualisation to the innovation translation theory as an augmented filter enabling an in-depth view of data. In Table 1, the conceptual framework incorporating the innovation translation 'moments' and ANT visualisation lens is presented along with the relevant excerpts from the literature.

**Table 1:**Conceptual Framework—Innovation Translation ‘Moments’ & ANT

	Moment 1 <i>Problematization</i>	Moment 2 <i>Interessement</i>	Moment 3 <i>Enrolment</i>	Moment 4 <i>Mobilisation</i>
Literature review (2005-13) see for example (Nagy et al., 2006; Fisher and Monahan, 2008; Cox, 2008; Cheng and Chai, 2012; Yao et al., 2012; Martinez-Perez, et al., 2012; Ustundag, 2013)	The key actors in hospitals seem to be technology vendors, systems integrators and administrators in hospitals. Users of the RFID system such as nurses and orderlies are ignored or forced to use the system. The moment is characterised by the acceptance of technology for tracking equipment, in order to reduce costs and improve efficiencies by hospital administrators and strategists.	As the literature revealed there is little evidence of getting users (nurses and orderlies) interested in RFID solution. Rather, driven by vendor and administrators of hospitals, other key actors are largely forced into the moment.	Enrolment into the RFID enablement network is forced or pushed through by vendors and administrators	The RFID solution has relatively low acceptance or rather scepticism from users. Since the users were sceptical the solution did not get promoted. Literature reveals that despite the scepticism, RFID as a technology was being promoted in the health sector via IT strategists and vendors.

The moment is characterised by the acceptance of technology for tracking equipment, in order to reduce costs and improve efficiencies by hospital administrators and strategists. As the literature revealed there is little evidence of getting users (nurses and orderlies) interested in RFID solution.

Rather, driven by vendor and administrators of hospitals, other key actors are largely forced into the moment. Enrolment into the RFID enablement network is forced or pushed through by vendors and administrators

The RFID solution has relatively low acceptance or rather scepticism from users. Since the users were sceptical the solution did not get promoted.

Literature reveals that despite the scepticism, RFID as a technology was being promoted in the health sector via IT strategists and vendors.

The moments of translation in table 1 reveal that despite the promotion by technology vendors and to an extent, ICT departments or people in hospitals, RFID technology is yet to gain momentum, due to use reluctance and non-acceptance (nurses or orderlies) within the Australian system. To a large extent, this solution is yet to gain wide acceptance in hospitals. It was evident in the ‘moments of translation’ that the key moments of interessement and enrolment are often a product of coercion or a top down approach, rather than mobilisation of allies in the environment. This factor made it imperative to study RFID implementation and its translation into the Australian health context.

### 3. Methodologies

We have chosen multiple case study methodology for this inquiry, using two landmark case sites (hospitals) in Australia. The choice was based on an Australian milieu being different for implementing RFID technology, as it touches every department in terms of ‘tracking’ ability, as well as people. The issue under investigation was innovation translation of RFID in a complex socio-technical context. This is not well understood, and the aim is to gain an in-depth understanding regarding ‘actors’, their opinions and relationships, and changes to their relationships and interactions. The research questions involved understanding how the actors in the network of translation negotiate with each other, to effect successful RFID translation. The process of negotiation and translation involves complex issues in hospital context: a chaotic environment where people continuously



negotiate with each other. Information relating to these issues of translation is best obtained through qualitative fieldwork (Creswell 2013). The lead researcher not only interviewed people in their settings, but also observed and recoded behaviours.

The choice of *case study* method is justified by Yin (2009) as an appropriate method when we tried to answer 'how' and 'why' questions, and we had less control over events being observed, and when the project was a contemporary phenomenon with a real life context. All of the above elements are present in this research that investigates RFID translation into Australian hospital contexts. Moreover, we have resorted to *purposeful sampling*, which is a concept where researchers select sites for the study because they can purposefully inform an understanding of the research problem and the central phenomenon in the study (Creswell 2013: 156). In this investigation, the determining criteria were that RFID implementation was partially or fully completed or at least initiated in the hospitals. We were able to find a pioneering, partially complete and a successful, fully completed situation, to draw cross-case comparisons later. As described in Creswell's (2013) compendium of data collection approaches of qualitative research, from six sources of evidence for cases as prescribed by Yin (2009), and drawing from many authors in the field, we have chosen observation, interviews and documents for data collection. For observation purposes, the principal researcher was a semi-participative observer (Case-1) and a non-participative observer (Case-2) and gathered field notes. A semi-structured approach was resorted to for interviews. Focus groups were also used for validation.

For analysing the interview data, concept mapping technique was used to reveal hidden patterns. These were then presented using ANT based information visuals, thus abstracting the data to the next level of information. This technique helped abstraction of real information without breaching the confidentiality and did not affect the sensitivity of the health environment.

#### 4. Findings (or Knowledge Abstraction) and Analytical Discussion

In this research investigation, the lead researcher went through a series of interviews and observations to collect relevant data. This data was transcribed and re-constituted in a '*movie script*' format with Acts/Scenes, actors and dialogue (See Figure 1 as an example). This unique method of presentation was then funnelled into an information visualisation through an ANT informed lens.

<i>Cameo:</i>	<i>What is your opinion on RFID being deployed here in ED?</i>
<i>Head-ED:</i>	<i>This is an area where there is frequent movement of equipment, people and other resources on a regular basis. In this chaotic environment, it is an expectation that equipment can be tracked immediately, and these are sterilised and the correct equipment. For example, often this department loses its wheelchairs while transporting patients, and over the last year, there has been a loss of 11 wheelchairs. Similarly, there is equipment that gets lost in laundry areas of wastebaskets due to the nature of emergency. There are no tracking or information systems in place at present, and collaborative care is therefore dependant on the goodwill of existing orderlies only (not RFID).</i>

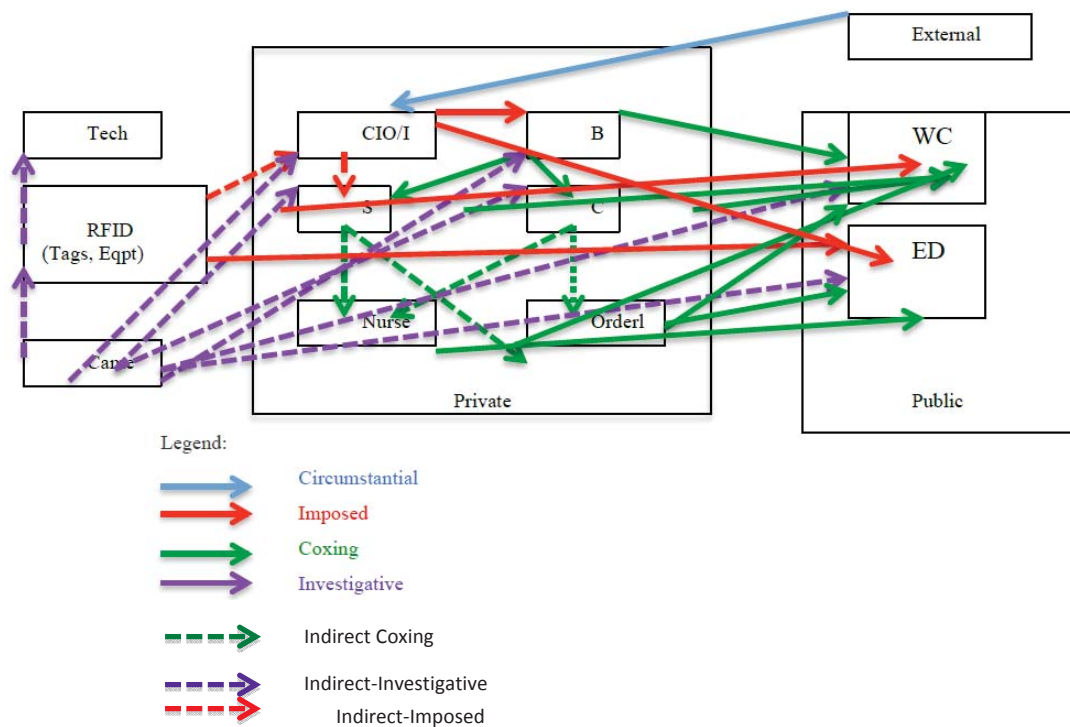
**Figure 1:** Case-1 (Scene-3): Meeting the Head of Emergency Department and Business Analyst

The lead researcher appears in the script as '*Cameo*' and RFID is the '*debut*' actor. The varied stakeholders such as technology implementers, CIOs, nurses and orderlies are the human actors; and there are also non-human actors such as artefacts, interfaces etc. Table 2 presents the entry and exit of actors for Case 1, in the black box of the hospital. And Figure 2 visualises the interplay between actors, formation of networks and the complexity of negotiation from Case 1.



**Table 2:** Case 1 – Actor-Network Relationships

Human Actors	Existed	Entered	Exited	Sustained
CIO (The Champion)	X			X
PM (Technology Company)				
External Company teams				
Head (SSD)	X			X
Nursing Unit Manager-Private	X			X
Orderlies (Private)	X			X
Nurses (Private)	X			X
SysAnalyst (Implementation Projects)	X			X
Business Analyst	X			X
Cameo		X	X	
Wards-in-Charge (Public)		X		X
Head (ED)		X		X
<b>NON HUMAN ACTORS</b>				
RFID Tags and Equipment		X		X
The Debut Actor				
RTLS Report		X		X
Web Based Clinical System	X			X
Bo-Beep Interface	X			X
Touch Screens	X			X
New Funding Group (Sirens)		X		X



**Figure 2:** Interplay of Actors, Network Formation and Relationships Negotiated (Case-1)

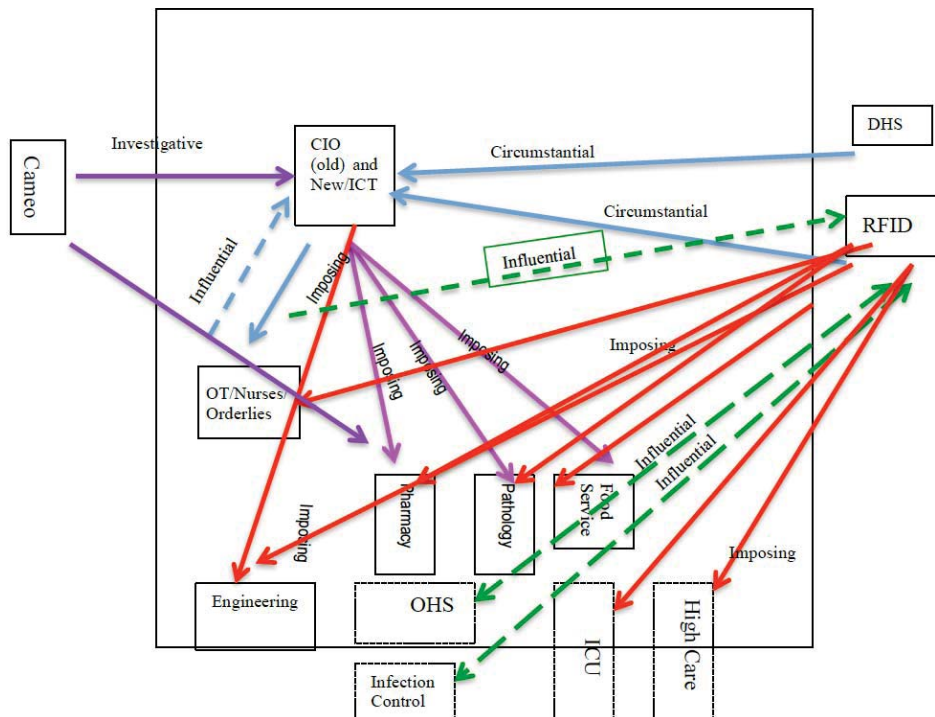
In this hospital, RFID deployment was initiated by the ICT department for the private wing and was abandoned by users time although the equipment remained in the hospital. The CIO and Business Analyst tried for 2 years pumping in discretionary funding into the cause of championing RFID, without any results. After 2 years, a nurse-in-charge of 'wound-care-wards' communicated with the head of the Emergency (also a nurse) in the Public Wing to re-initiate RFID in the hospital. Subsequently, RFID became the buzzword again, as they

communicated in a different language to the operations people—namely orderlies—in the hospital. This triggered the redeployment funded also externally within the public hospital. It is interesting to note that the hospital management, the CIO and ICT department, other department heads, clinicians and orderlies heeded the nurses' communication. It only took the wound-care-ward in charge a few months before the 'stage was reset' to deploy RFID again.

Subsequently, in Table 3, we present the actor-network relationships in Case-2, followed by Figure 4 visualising the actor-network relationships formed and their complexity.

**Table 3:** Case 2- Actor-Network Relationships

Actors	Existed (in the blackbox)	Entered	Exited	Sustained/Transformed
DHS/ PPPs		X		X
Symposium Article		X		X
RFID (tags, eqpt) (External Company)		X		X
RFID Results		X		X
RFID Maps		X		X
Infection Control	X			X
Cameo		X	X	
Nursing Staff	X			
Orderlies	X			
Pharmacy	X	X		X
Pathology	X	X		X
Food Services	X	X		X
Engineering	X	X		X
CIO (new)	X	X		X
OHS	X			X
ICU	X			X
High Care	X			X



**Figure 4:** Actor-Network Relationships (Case-2)

In this hospital, RFID had been introduced by an orthopaedic nurse and the technology itself had evolved taking on many facets and it was found useful by varied departments, as it translated into many formats such as temperature tags, asset tags, and alarms. It also enabled a web of relationships subtly and indirectly, via its champion – the nurse.

## **5. Conclusions and Implications**

The network of actors and their relationships is extremely complex and is key to operations in a hospital. In this investigation, in Case-1, both orderlies and nurses felt disempowered by technology introduction. The nurses did not endorse imposing a technology that disrupted the workflow. And it had taken 3 years and a nurse-in-charge to re-deploy the technology that was then accepted with ease by all stakeholders including clinicians. In Case-2, the introduction by a nurse enabled successful translation of RFID. The findings revealed a silent web of relationships between the key actors in hospitals, in relation to promoting RFID technology. The constant communication flow between orderlies to orderlies, nurses to orderlies, and nurse-to-nurse across the private and public areas of Case-1 and nurse–nurse, nurse–clinician and nurse–ICT relationships, which are not clearly visible at the onset, is indeed the most powerful social factor for RFID implementation.

Additionally, expert opinion/validation from industry sector indicate that in Australian hospitals, nurse is the powerful and influential factor in technology translation. For example, the ICT department feels imposed upon by medical directors, but if the nurse is the person raising the issue, they will accept take it on board and enable it. Doctors do not question nurses neither do the patient care orderlies. Nurse happens to be the *lynchpin* in Australian hospitals. This knowledge abstraction is significant from the Australian perspective, which would not have been possible without the interpretive stance of this investigation, visualising the knowledge through an ANT informed lens and eliciting the key social factors in translation further validated by industry experts. The finding is of significant value to large hospitals on the verge of RFID deployment. Conversely, the research extended Innovation Translation theory framework and augmented the field of ANT by visualisation techniques. This value addition is significant implication for academia as it added to the body of knowledge that is currently rather limited in the field of Health Informatics within Australia.

## **Acknowledgement**

As a PhD candidate at Victoria University - College of Business, the lead author, Chandana Unnithan, wishes to thank all the staff for their support in completing this project successfully.

## **References**

- Callon, M. (1986) Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay, in J. Law, *Power, action and belief: a new sociology of knowledge?* London, Routledge, pp.196-223.
- Cangialosi, A., Monaly, J.E. and Yang, S.C. (2007) Leveraging RFID in hospitals: Patient lifecycle and mobility perspectives, *Communications Magazine*, IEEE, Volume 45(9), pp. 18–23.
- Coustasse, A., Tomblin, S. and Slack, C. (2013) A review of Radio Frequency Identification Technologies and Impacts on the Hospital Supply Chain: 2002-2012, Proceedings of Academic and Business Research Institute (AABRI) International Conference, MMM Track, January 3-5, Orlando, Florida.
- Creswell, J.W. (2013) *Research Design: qualitative, quantitative and mixed method approaches*, 2<sup>nd</sup> Ed, Sage Publications, Thousand Oaks, California.
- Davis, F. (1986) *A technology acceptance model for empirically testing new end user information systems: Theory and Results*, MIT, Boston.
- Degaspari, J.J. (2011) Keeping track. Barcodes and RFID tags make inroads in hospitals. *Healthcare informatics: the business magazine for information and communication systems*, 28(3), pp. 44–47.
- Duckett, S.J. and Willcox, S. (2011) *The Australian health care system*, Edition 4, Melbourne: Oxford University Press.
- Dunlevy, S. (2013) Outrage as eHealth record sign-up squads hit Australian hospital patients in bid to boost numbers, News.com.au, Retrieved April 30 2013 from <http://www.news.com.au/national-news/outrage-as-ehealth-record-sign-up-squads-hit-australian-hospital-patients-in-bid-to-boost-numbers/story-fncynjr2-1226619874616>
- Greenhalgh, T. and Stones, R. (2010) Theorising big IT programmes in healthcare: strong structuration theory meets actor–network theory. *Social Science & Medicine*, 70(9), pp. 1285–1294.
- Ho, G. (2012) Can technology help overcome Australia's healthcare challenges? Australian Broadcasting Corporation (ABC) - Technology and Games, 27 April 2012, Retrieved in January 2014 from <http://www.abc.net.au/technology/articles/2012/04/27/3490209.htm>
- Latour, B. (1986) The power of association, Power, Action and Belief – a new Sociology of Knowledge, Sociological Review Monograph 32, Law J (Ed), Routledge and Kegan Paul, London, pp. 264-280
- Novak, J. and Judah, A. (2011) *Towards a health productivity reform agenda for Australia*, South Melbourne: Australian Centre for Health Research.

- Parker, C.M., Wafula, E.N., Swatman, P. and Swatman, P.M.C. (1994) Information Systems Research Methods: The Technology Transfer Problem, Proceedings of the Australasian Conference on Information Systems (ACIS), Australia.
- Prgoment, M., Georgiou, A. and Westbrook, J.I. (2009) The impact of mobile handheld technology on hospital physicians work practices and patient care: A systematic review, American Medical Information Association, Nov–Dec, Volume 16(6), pp. 792–801.
- Privacy Act of Australia (1988) <http://www.privacy.gov.au/law/act>
- Privacy.Gov (2013) State and Territory Laws, Office of the Australian Privacy Commissioner, Australia, <http://www.privacy.gov.au/law/states>
- Productivity Commission (2005) Impacts of advances in medical technology in Australia, *Productivity Commission Research Report*, Melbourne Australia <http://www.pc.gov.au/study/medicaltechnology/finalreport/medicaltechnology.pdf>
- Rogers EM (2003) Diffusion of Innovations, Free Press, New York.
- Tatnall, A. (2011) *Information Systems Research, Technological Innovation and Actor–Network Theory*, Heidelberg Press, Victoria, Australia.
- Whetton, S. and Georgiou, A. (2010) Conceptual Challenges for Advancing the Socio-Technical Underpinnings of Health Informatics, *The Open Medical Informatics Journal*, Volume 4, 221–224
- Yao, W., Chu Chao-Hsien. and Li Zang. (2011) Leveraging complex event processing for smart hospitals using RFID, *Journal of Network and Computer Applications*, 34(3), May, pp.799–810
- Yao. W., Chu Chao-Hsien. and Li Zang (2012) The Adoption and Implementation of RFID technologies in Healthcare: A literature review, *Journal of Medical Systems*, 36, pp. 3507–3525, Springer Publications, USA.



# **Masters Research paper**





# Finding Patterns in Medical Ward Data Using Rough Sets

Anwar Alenezi and Julia Johnson

Department of Mathematics and Computer Science

Laurentian University, Sudbury, Canada

[ax\\_alenezi@laurentian.ca](mailto:ax_alenezi@laurentian.ca)

[jjohnson@cs.laurentian.ca](mailto:jjohnson@cs.laurentian.ca)

**Abstract:** Data have been obtained from a hospital in Saudi Arabia. In this project, we are discovering patterns an experimental tool called Rough Set Graphic User Interface (RSGUI). Several algorithms are available in RSGUI, each of which is based in Rough Set theory. Our objective is to find short meaningful predictive rules. First we find a minimum set of attributes that fully characterize the data. Some of the rules generated from this minimum set of attributes were obvious and therefore uninteresting. Others were surprising and therefore interesting. RSGUI allows fine-tuning of rules making it possible to analyze information about the rules and compare the rules resulting from different algorithms. Usual measures of strength of the rule, such as length of the rule, certainty and coverage were considered. In addition, a measure of interestingness of the rules has been developed based on questionnaires administered to human subjects.

**Keywords:** Rough Set, Data Mining, Health Analytics, Uncertainty, and Decision Making

---

## 1. Introduction

A great many decision making systems are in use to help doctors care for their patients, but not so to help hospital administrators care for their hospitals. Indeed, analytic approaches to help improve the healthcare system itself are a recent development (Saxena & Srinivasan, 2013). Our aim is to employ knowledge discovered in health information to help improve the quality of healthcare delivery and the effectiveness of healthcare managers. An innovative solution has been developed based on an analytic and knowledge management approach to decision making.

### 1.1 RSGUI

Rough Set Graphic User Interface (RSGUI) (Campeau & Johnson, 2005) is a software system providing different algorithms appropriate for decision making (Johnson & Johnson, 2008). The different algorithms that RSGUI features are RS1 that was originally developed by Wong and Ziarko (S. K. M. Wong & Ziarko, 1986) (S.K.M. Wong & Ziarko, 1986) and improved upon in the following years (LEM, LEM2 (Grzymala-busse, 2010)), an Inductive Learning Algorithm (ILA) (Tolun & Abu-Soud, 2007), and Rough Set Reverse Prediction Algorithm (RSRPA) (Johnson & Campeau, 2005). The system has been developed by Laurentian University Computer Science students beginning in 2005 and continues to evolve as students make enhancements. RSGUI serves as the framework for this research.

## 2. Rough Sets Theory

Rough set theory is concerned with uncertainty and imprecise data. It can be used to extract meaningful rules from datasets. It attempts to approximate a given concept based on a minimum amount of decision rules. Rough set theory has very important concepts such as lower approximation and upper approximation (Rissino & Lambert-torres, 2009). Lower approximation consists of all elements that surely belong to a concept described in the dataset. On the other hand, upper approximation consists of all elements that possibly belong to the concept. Therefore, the difference between both approximations (upper - lower) is called the boundary region. This region contains cases on the boundary that cannot be precisely classified. Let  $[x]_p$  denote, the rows of a table that cannot be distinguished one from the other by means of their properties  $p$ . Then, the lower approximation to concept  $X$  is defined as follows:

$$\underline{RX} = \{x \mid [x]_p \subseteq X\}$$

The lower approximation is also called the positive region and it is the union of all equivalence classes  $[x]_p$ .

The upper approximation is defined as follows:

$$\overline{RX} = \{x \mid [x]_p \cap X \neq \emptyset\}$$

The upper approximation is also called the negative region. The boundary region is defined as follows:

$$\text{Boundary}(X) = \overline{RX} - \underline{RX}$$

We can say that the set is rough if its boundary set is nonempty. Otherwise the set is exact (Rissino & Lambert-torres, 2009).

### **3. Nature of the data**

Data have been obtained from a general hospital in Saudi Arabia. The hospital agreed to let us use the data of the summer of 2013. The statistical data provided for this research project is unidentifiable. Also the Research Ethics Board at Laurentian University has approved use of these data in this research. Data have been obtained for three months period.

All experiments reported here were done upon the Female Medical Ward (FMW) data table. The hospital has 10 wards, so the process that gives the results of FMW can be applied to the rest of the wards also.

#### **3.1 Methodology**

The medical ward data may have useful implicit rules that are not immediately obvious by viewing the table directly. To get those rules, careful selection of important or interesting decision attributes is needed. In this case, the first decision attribute that will be checked is occupancy to see what could make the ward busy in order to limit the busyness if it is causing problems. For example, sometimes emergency doctors unnecessarily admit patients, when a medicine may be enough to treat their condition. The second decision might be death either to see if the occupancy caused some deaths or to know at least that the deaths cannot be controlled by the staff in the ward.

After testing the data by RSGUI, we are going to evaluate the rules that RSGUI generated. Criteria by which to evaluate the rules generated will be describe in more details, in section 5. The most important matter is to find the useful and interesting rules from the data and order them from the most interesting to least interesting. Many questions may come to minds for the hospital director and the manager of Computer and Statistical department. For example, what can make the ward busy? Is ward occupancy a factor contributing to staff negligence? Does staff negligence cause deaths in a ward?

The hospital sends an inpatient statistical report to the Ministry of Health Care daily. The reasons that the Ministry has asked for reports, are to analyze inpatient statistics, to verify if there is enough space, to see what can cause the busyness, and which of those previous reasons could have caused deaths in the ward? The Ministry could expand the ward with more rooms, beds, or employ more staff and doctors to at least limit busyness ward, to remedy the situation. Another reason that may cause deaths or busyness, is that the staff or doctors may not be qualified enough to help their patients, so the mentioned remedial actions be done to deliver a high quality patient care.

Rough Set1 (RS1) is a data mining algorithm for decision making based on incomplete, inconsistent, imprecise and vague data. Most Healthcare data have these properties. Soft computing techniques of which rough sets are one along with fuzzy sets aims to be as precise as possible about imprecision. New knowledge in the form of predictive rules is automatically discovered and a measure of the plausibility of inferences is given. Rough set and soft computing in general are data analytics approaches because implicit, previously unknown and hidden information is brought to the surface. Knowledge discovered from the ward data will be useful for the Ministry of Health Care for policy making, and for all levels of hospital management, for improving patient safety and satisfaction and for improving hospital productivity in general.

### **4. Discretization**

“Discretization is the process of turning continuous values into discrete intervals” (Kotsiantis & Kanellopoulos, 2006). The data used in this thesis need to be discretized because the data comprise numerical and continuous values. There is an abundance of discretization algorithms(Yang & Webb, 2002)(Siekmann, Hartmanis, & Leeuwen, 1998)(Kotsiantis & Kanellopoulos, 2006)(Wu, 2004). Any one of these could have been used. Discretization for obtaining discrete values from numerical data is a well understood problem. To get the job done, I looked at natural clustering of the data by visual inspection and also used equal frequency binning discretization method(Frank, 2005). Table 1 is a partial sample of one month ward data before discretization.

**Table 1:** Sample of five days of Female Medical Ward (FMW) data

Day	Remain Patient	New Admission	Death	Discharge	Occupancy
1	18	1	0	7	12
2	12	2	0	1	13
3	13	6	0	3	16
4	16	6	0	4	18
5	18	5	0	6	17

The discretization method will be described next using the first column (Remain patient column) for illustration and the same discretization process goes for all other columns, Table 1 is a sample of 30 rows and 6 columns. There are 5 columns in total on the above table that need to be discretized.

Grab all distinct numbers in the first column. Here we have the minimum number is 9 and maximum is 24. Suppose we want to divide the first column into 3 bins. That means in this example each bin has 4 elements. 9, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 24

The boundary values of the bins are as follows:

$$(14+15)/2 = 14.5, \quad (18+19)/2 = 18.5$$

So, after that the value ranges will be as follows:

Bin 1: [0, 14.5] → LOW

Bin 2: [14.5, 18.5] → NORMAL

Bin 3: [18.5, +∞] → LARGE

Table 2 shows the sample of one month ward data from Table 1 discretized to be as Table 2 which that now the column entries are all soft values.

**Table 2:** Discretized table for Female Medical Ward data for one month

Day	Remain Patient	New Admission	Death	Discharge	Occupancy
1	NORMAL	LOW	GOOD	LARGE	NOT_BUSY
2	LOW	LOW	GOOD	LOW	NOT_BUSY
3	LOW	LARGE	GOOD	LOW	NORMAL
4	NORMAL	LARGE	GOOD	NORMAL	NORMAL
5	NORMAL	NORMAL	GOOD	NORMAL	NORMAL

#### 4.1 Experiments to determine interesting rules

Table 3 is a summary of all experiments that have done in RSGUI using Female Medical Ward data. RS1 and ILA algorithms have been applied upon the data for one month and three months. So the Table 3 shows the attributes that have been used in the experiments, how many rows were considered, how many columns were considered, and how many rules we got from every experiment.

**Table 3:** Summary of the FMW data in RSGUI

Algorithm	Conditions Attributes	Decision Attributes	Rows Considered	Columns Considered	Number of rules
RS1	remain, admitted, death, discharges	occupancy	29	5	13
ILA	remain, admitted, death, discharges	occupancy	29	5	14
RS1	remain, admitted, death, discharges	occupancy	77	5	30
ILA	remain, admitted, death, discharges	occupancy	77	5	24
RS1	remain, admitted, death, discharges	occupancy	89	5	39
RS1	remain, admitted, discharges, occupancy	death	89	5	37

## 5. Interestingness measures

User uses the RSGUI to discover patterns in data. To get a best result from these data, it needs to extract interesting rules. For the rule to be interesting, it needs to be useful and eventually understandable. Interestingness measures should be found by two parts: objective and subjective.

Interestingness measures are for selecting and ranking patterns according to how interesting they are from the user's perspective. A good survey of interestingness measures can be found in (Geng & Hamilton, 2006). In that work, criteria to determine interest of patterns are the following:

*Conciseness*: a rule can be concise if it has few attributes relative to the others.

*Generality/ Coverage*: a rule can be general or coverage if it covers a large subset of the data.

*Reliability*: a rule can be reliable if it is strong as measured by some metric.

*Peculiarity*: a rule is peculiar if it is unusual and unknown beforehand to the user.

*Diversity*: a rule is diverse if its attributes differ from each other.

*Novelty*: a rule is novel if the user does not know it beforehand and cannot produce it from known rules.

*Surprisingness*: a rule is surprising if it is unexpected and contradicts user knowledge.

*Utility*: a rule has utility if it is useful and helps a person to reach a goal.

*Applicability*: a rule is applicable if it can be applied in the future in same domain.

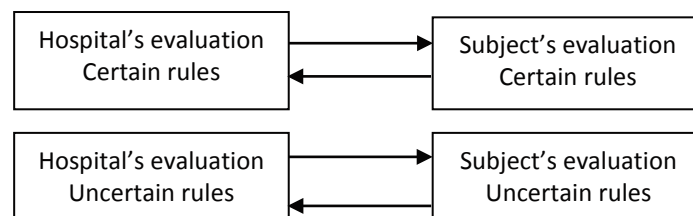
In (Greco, Roman, & Szc, 2012) alternate approaches to normalization of measures are provided, which means they propose methods to make different measures comparable. A rule generated may be good on some and bad on others, so developing a formula for taking into consideration all the evaluation criteria is needed to get a good result. The formula that I developed uses ranking from -4 to +4 for each criterion and adding up the positive and negative scores yields to obtain one weight for the rule. The rankings for each individual criterion are as follows:

- +4 Reliability
- +3 Utility
- +2 Conciseness
- +1 Generality/ Coverage
- 0 Novelty
- 1 Applicability
- 2 Diversity
- 3 Surprisingness
- 4 Peculiarity

### Evaluation

Two experiments have done, one upon one of the hospital administrators and the other upon Laurentian University students. In order to get a best evaluation of the outcomes, we asked one of the hospital administrators to evaluate the rules from most interesting to least interesting according to the interestingness criteria.

The difference between the hospital administrator evaluation and each subject evaluation has made to check how far the subject opinion is from the hospital evaluation. Figure 3 shows the degree of departure from both evaluations



**Figure 1:** Degree of departure from the hospital evaluation

We assume that difference between student evaluation of certain rules and hospital administrator's evaluation gives error of student perspective. For finding degree of precision of uncertain rules, the error on the part of the students is subtracted from their evaluation of uncertain rules.

### 5.1 Occupancy decision attribute evaluation rules

RSGUI has generated 39 rules from the FMW data with remain, admitted, discharged, and death as condition attributes and occupancy as a decision attribute. Some rules were removed because they were not interesting as evaluated by the hospital and because coverage was low as computed by RS1 algorithm. At the end, we came up with 17 rules to evaluate, First 5 rules predict busy ward, middle 7 rules predict normal ward, and last 5 rules predict not busy ward. The 17 occupancy decision rules are as follows:

1. (remain := LARGE) AND (discharged := LOW) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
2. (remain := LARGE) AND (discharged := NORMAL) AND (admitted := LARGE) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
3. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := NONE) ---> (occupancy := BUSY)[certainty = 1.0][coverage = 2/89]
4. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := 1) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 1/89]
5. (remain := NORMAL) AND (discharged := LOW) AND (admitted := NORMAL) AND (death := NONE) -----> (occupancy := BUSY)[certainty = 0.2727272727272727][coverage = 11/89]
6. (discharged := LARGE) AND (death := NONE) AND (remain := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]
7. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 5/89]
8. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 3/89]
9. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.9333333333333333][coverage = 15/89]
10. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.7272727272727273][coverage = 11/89]
11. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.5][coverage = 6/89]
12. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LOW) -----> (occupancy := NORMAL)[certainty = 0.8333333333333334][coverage = 6/89]
13. (admitted := LOW) AND (remain := LOW) -----> (occupancy := NOT\_BUSY)[certainty = 1.0][coverage = 5/89]
14. (admitted := LOW) AND (remain := NORMAL) AND (discharged := LARGE) -----> (occupancy := NOT\_BUSY)[certainty = 1.0][coverage = 2/89]
15. (admitted := NORMAL) AND (remain := NORMAL) AND (discharged := LARGE) AND (death := NONE) -----> (occupancy := NOT\_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
16. (admitted := LOW) AND (remain := NORMAL) AND (discharged := NORMAL) AND (death := NONE) -----> (occupancy := NOT\_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
17. (admitted := NORMAL) AND (remain := LOW) AND (discharged := LOW) AND (death := NONE) -----> (occupancy := NOT\_BUSY)[certainty = 0.5][coverage = 6/89]

Evaluation Tools:

The 17 rules sent to the hospital administrator and subjects with an empty evaluation table same as Table 6. A description of the evaluation criteria was given to hospital administrator and subjects. At the top of the rules was a description of the ward domain and the fields named in the rules to help them evaluate the rules.

Hospital administrator evaluation results for occupancy decision attribute shown on Table 6.

**Table 4:** Hospital administrator evaluations for occupancy decision attribute

Rule#	Reliability	Utility	Conciseness	Generality	Novelty	Applicability	Diversity	Surprisingness	Peculiarity	Sum of the rule weight
1	✓	✓	✓	✓		✓	✓			7
2		✓		✓		✓				3
3	✓	✓				✓	✓			4
4	✓	✓				✓	✓			4
5		✓		✓				✓		1
6			✓		✓				✓	-2
7				✓	✓			✓		-2
8					✓			✓		-3
9	✓	✓		✓		✓				7
10				✓				✓		-2
11		✓		✓		✓				3
12		✓		✓		✓				3
13	✓	✓	✓	✓		✓				9
14		✓	✓			✓	✓			2
15		✓			✓	✓				2
16		✓				✓				2
17				✓	✓			✓	✓	-6

Each criterion was associated with a weight as described in section 5 to reflect the importance of that particular criterion. Recall that the weights range from -4 to +4. When the subject checked the box under a criterion the weight associated with it was accumulated in a running sum. The final sum gives the overall evaluation of the rule. For example:

Frist row: the reliability, utility, conciseness, generality, applicability, and diversity have checked. The final sum of the first row is computed as follows:

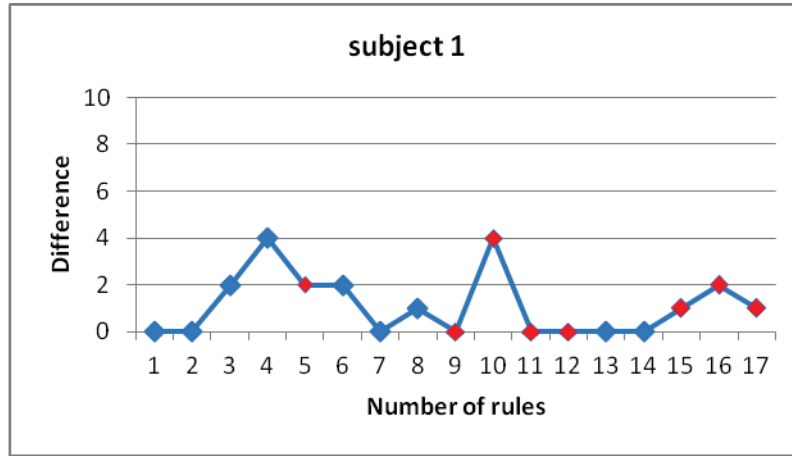
4 + 3 + 2 + 1 – 1 – 2 respectively. The final sum for these numbers is 7. This process is applied to all 17 rules to come with the results on the last right column on Table 6.

Order of the result after the evaluation from the most interesting, which is the higher number to least interesting, which is the smallest number according to the sum of the rule weight on Table 6 would be as follows:

1. (admitted := LOW) AND (remain := LOW) -----> (occupancy := NOT\_BUSY)[certainty = 1.0][coverage = 5/89]
2. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.9333333333333333][coverage = 15/89]
3. (remain := LARGE) AND (discharged := LOW) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
4. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := NONE) ---> (occupancy := BUSY)[certainty = 1.0][coverage = 2/89]
5. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := 1) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 1/89]
6. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LOW) -----> (occupancy := NORMAL)[certainty = 0.8333333333333334][coverage = 6/89]
7. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.5][coverage = 6/89]
8. (remain := LARGE) AND (discharged := NORMAL) AND (admitted := LARGE) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
9. (admitted := LOW) AND (remain := NORMAL) AND (discharged := LARGE) -----> (occupancy := NOT\_BUSY)[certainty = 1.0][coverage = 2/89]
10. (admitted := LOW) AND (remain := NORMAL) AND (discharged := NORMAL) AND (death := NONE) -----> (occupancy := NOT\_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]

11. (admitted := NORMAL) AND (remain := NORMAL) AND (discharged := LARGE) AND (death := NONE) -----> (occupancy := NOT\_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
12. (remain := NORMAL) AND (discharged := LOW) AND (admitted := NORMAL) AND (death := NONE) -----> (occupancy := BUSY)[certainty = 0.2727272727272727][coverage = 11/89]
13. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 5/89]
14. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.7272727272727273][coverage = 11/89]
15. (discharged := LARGE) AND (death := NONE) AND (remain := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]
16. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 3/89]
17. (admitted := NORMAL) AND (remain := LOW) AND (discharged := LOW) AND (death := NONE) -----> (occupancy := NOT\_BUSY)[certainty = 0.5][coverage = 6/89]

Also the evaluation has been done upon 24 Laurentian University students, 4 of them are studying in the medical school at Laurentian University. Figure 6 is one example of the degree of departure of subject 1 evaluation from the hospital evaluation.



**Figure 2:** Difference between subject evaluation and hospital evaluation

The rules (5, 9, 10, 11, 12, 15, 16, and 17) are uncertain rules. We can see that the difference of subjects evaluation were close enough to the hospital administrator evaluation. Once the difference is close to 0, which mean the subject evaluation is close enough to the hospital administrator evaluation or even same.

The difference computed as follows:

$$|a| - |b| = |d| \quad (\text{Figure 6})$$

Where  $a$  is the sum of hospital administrator evaluation for one rule,  $b$  is the sum of subject evaluation for same that rule, and  $d$  is the difference between both evaluation. Once there is no matter if the difference on positive or negative side, we take the absolute value of the  $d = |d|$ . Few subject evaluations were ignored because their differences were close to 10 and far away from 0.

## 5.2 Death decision attribute evaluation rules

Subsection 6.2 is same process of subsection 6.1. The rules that RS1 generated were 37, some were removed because they were not interesting as evaluated by the hospital and because coverage was low by computed by RS1 algorithm. At the end, we came up with 15 rules to evaluate, First 3 rules predict 2 deaths in the ward, middle 5 rules predict 1 death in the ward, and last 7 rules predict no deaths in the ward. The 15 death decision rules are as follow:

1. (remain := LARGE) AND (admitted := NORMAL) AND (occupancy := NOT\_BUSY) ----> (death := 2)[certainty = 1.0][coverage = 1/89]
2. (remain := NORMAL) AND (admitted := LARGE) AND (occupancy := NORMAL) AND (discharged := LOW) ----> (death := 2)[certainty = 1.0][coverage = 1/89]



3. (remain := LARGE) AND (admitted := LOW) AND (occupancy := NORMAL) AND (discharged := NORMAL) ---> (death := 2)[certainty = 0.25][coverage = 4/89]
4. (discharged := LARGE) AND (admitted := NORMAL) AND (occupancy := BUSY) --->(death := 1)[certainty = 1.0][coverage = 1/89]
5. (discharged := NORMAL) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := LARGE) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
6. (discharged := NORMAL) AND (admitted := LOW) AND (occupancy := NORMAL) AND (remain := LARGE) ---> (death := 1)[certainty = 0.25][coverage = 4/89]
7. (discharged := LOW) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
8. (discharged := LOW) AND (admitted := NORMAL) AND (occupancy := NOT\_BUSY) AND (remain := LOW) ---> (death := 1)[certainty = 0.25][coverage = 4/89]
9. (discharged := LOW) AND (admitted := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 12/89]
10. (discharged := NORMAL) AND (admitted := LARGE) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 5/89]
11. (discharged := NORMAL) AND (admitted := NORMAL) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 15/89]
12. (discharged := NORMAL) AND (admitted := LARGE) AND (remain := LARGE) AND (occupancy := BUSY) ---> (death := NONE)[certainty = 0.6666666666666666][coverage = 3/89]
13. (discharged := LOW) AND (admitted := NORMAL) AND (remain := NORMAL) AND (occupancy := NORMAL) --> (death := NONE)[certainty = 0.8888888888888888][coverage = 9/89]
14. (discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.75][coverage = 4/89]
15. (discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NOT\_BUSY) ---> (death := NONE)[certainty = 0.75][coverage = 4/89]

Hospital administrator evaluation result for death decision attribute shown on Table 5.

**Table 5:** Hospital administrator evaluations for death decision attribute

Rule#	Reliability	Utility	Conciseness	Generality	Novelty	Applicability	Diversity	Surprisingness	Peculiarity	Sum of the rule weight
1			✓				✓	✓	✓	-7
2					✓			✓		-3
3		✓		✓			✓			2
4		✓	✓			✓				4
5	✓	✓				✓				6
6		✓				✓				2
7	✓	✓				✓	✓			4
8				✓	✓					1
9		✓	✓	✓	✓					6
10		✓	✓	✓						6
11	✓	✓	✓	✓		✓				9
12						✓		✓		-4
13		✓		✓						4
14		✓		✓						4
15	✓	✓		✓						8

Death evaluation rules ordered from most interesting to least interesting according to the weight criteria from the hospital administration evaluation Table 5 are as follows:

1. (discharged := NORMAL) AND (admitted := NORMAL) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 15/89]
2. (discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NOT\_BUSY) -> (death := NONE)[certainty = 0.75][coverage = 4/89]
3. (discharged := LOW) AND (admitted := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 12/89]
4. (discharged := NORMAL) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := LARGE) --> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
5. (discharged := NORMAL) AND (admitted := LARGE) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 5/89]
6. (discharged := LARGE) AND (admitted := NORMAL) AND (occupancy := BUSY) --->(death := 1)[certainty = 1.0][coverage = 1/89]
7. (discharged := LOW) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
8. (discharged := LOW) AND (admitted := NORMAL) AND (remain := NORMAL) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.8888888888888888][coverage = 9/89]
9. (discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.75][coverage = 4/89]
10. (remain := LARGE) AND (admitted := LOW) AND (occupancy := NORMAL) AND (discharged := NORMAL) ---> (death := 2)[certainty = 0.25][coverage = 4/89]
11. (discharged := NORMAL) AND (admitted := LOW) AND (occupancy := NORMAL) AND (remain := LARGE) ---> (death := 1)[certainty = 0.25][coverage = 4/89]
12. (discharged := LOW) AND (admitted := NORMAL) AND (occupancy := NOT\_BUSY) AND (remain := LOW) ---> (death := 1)[certainty = 0.25][coverage = 4/89]
13. (remain := NORMAL) AND (admitted := LARGE) AND (occupancy := NORMAL) AND (discharged := LOW) ---> (death := 2)[certainty = 1.0][coverage = 1/89]
14. (discharged := NORMAL) AND (admitted := LARGE) AND (remain := LARGE) AND (occupancy := BUSY) -> (death := NONE)[certainty = 0.6666666666666666][coverage = 3/89]
15. (remain := LARGE) AND (admitted := NORMAL) AND (occupancy := NOT\_BUSY) ---> (death := 2)[certainty = 1.0][coverage = 1/89]

Regarding to the results of most interesting to least interesting, we can see that if the discharges were low in the ward, it may cause a busy ward. Therefore, keeping patients long time in a ward as a needless or a patient option cause some problems. Also some private hospitals do not discharge their patient once the patients are not sick or getting better. They are trying to admit patients as long as they can to charge patients more money. They are illusory patients that they need more stay to take care.

Knowledge discovered in health information has been employed to help improve the quality of healthcare delivery and the effectiveness of healthcare managers. An innovative solution has been developed based on an analytic and knowledge management approach to decision making.

## 6. Conclusion

Upon examination of data about the activity (admissions, discharges, remaining patients, and deaths) on medical wards, a number of questions come to mind, such as, what conditions of the ward tend to increase patient death? This project was an experiment to find out how to use data mining tools based on rough set theory to uncover hidden relationships in medical ward data. The objective was to approximate a given concept such as "Hospital ward busy" or "Death on a ward" in terms of information about the ward. The power of rough set theory lies in its ability to abstract away information that is not essential leaving a minimum set of attributes for predicting interesting outcomes. The method was applied in reverse as a means of understanding essential factors that will lead to a desired outcome.

Methodical experimentation with RSGUI on the medical ward data obtained from the hospital has resulted in the selection of an interesting set of rules as measured by parameters to describe rules. Those parameters allowed rules to be calibrated until the interesting ones have been discovered

Many applications use commercially available rough sets based tools that cannot be modified, but RSGUI allows modification of its own code because it has been developed by Laurentian University students over the years and it is open to further development. A new version of RSGUI has been embedded in a web interface for accessing and analyzing medical ward data.

## References

- Campeau, P. (2005). Predicting the Most Favorable Behavior of Artificial Objects using Rough Sets. Laurentian University, Sudbury, Canada.
- Frank, I. H. W. and E. (2005). Data mining: practical machine learning tools and techniques (2nd ed.). San Francisco: Morgan Kaufmann.
- Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining. *ACM Computing Surveys*, Vol 38, No. 3. pp 1-32.
- Greco, S., Roman, S., & Szcz, I. (2012). Properties of rule interestingness measures and alternative approaches to normalization of measures. *Information Sciences: An International Journal*, 216, 1–16.
- Grzymala-busse, J. W. (2010). A Local Version of the MLEM2 Algorithm for Rule Induction 2 . Blocks of Attribute-Value Pairs. *Fundamenta Informaticae*, 100, 99–116. doi:10.3233/FI-2010-265
- Johnson, J., & Campeau, P. (2005). Reverse Prediction, 88–97.
- Johnson, J., & Johnson, G. (2008). RSGUI with Reverse Prediction Algorithm, 287–306.
- Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization Techniques : A recent survey, 32(1), 47–58.
- Rissino, S., & Lambert-torres, G. (2009). Rough Set Theory – Fundamental Concepts , Principals , Data Extraction , and Applications. *Data Mining and Knowledge Discovery in Real Life Applications*, (February), 35–58.
- Saxena, R. N., & Srinivasan, A. (2013). Business analytics: a practitioner’s guide. New York: International Series in Operations Research & Management Science.
- Siekmann, J., Hartmanis, J., & Leeuwen, J. Van. (1998). *Lecture Notes in Artificial Intelligence*.
- Tolun, M., & Abu-Soud, S. (2007). An Inductive Learning Algorithm for Production Rule Discovery. Ankara, Turkey, Department of Computer Engineering Middle East Technical University, 1–19.
- Wong, S. K. M., & Ziarko, W. (1986). A machine learning approach to information retrieval. In: *Proceedings of the 9th Int. Conf. on R & D in Information Retrieval*, 228–233.
- Wong, S. K. M., & Ziarko, W. (1986). On learning and evaluation of decision rules in the context of rough sets. *Proceedings of the ACM SIGART International Symposium on Methodologies for Intelligent Systems*, 308–324.
- Wu, C. and M. L. and Z. H. and Y. Z. and Y. Y. (2004). Discretization Algorithms of Rough Sets Using Clustering. *Robotics and Biomimetics*, 2004. ROBIO 2004. IEEE International Conference on, 955–960.
- Yang, Y., & Webb, G. I. (2002). A Comparative Study of Discretization Methods for Naive-Bayes Classifiers. *Proceedings of PKAW*, 159–173.

# **Non Academic Papers**



# Statistical Validation of a new Python-Based Military Workforce Simulation Model

Stephen Okazawa, Patricia Moorhead, Abe Jesion and Stan Isbrandt

Defence Research and Development Canada, Ottawa, Canada

[stephen.okazawa@drdc-rddc.gc.ca](mailto:stephen.okazawa@drdc-rddc.gc.ca)

[patricia.moorhead@drdc-rddc.gc.ca](mailto:patricia.moorhead@drdc-rddc.gc.ca)

[abe.jesion@forces.gc.ca](mailto:abe.jesion@forces.gc.ca)

[stan.isbrandt@drdc-rddc.gc.ca](mailto:stan.isbrandt@drdc-rddc.gc.ca)

**Abstract:** The Canadian Department of National Defence uses military workforce modelling and simulation to inform the decision-making process with regard to the management of military personnel. Defence Research and Development Canada (DRDC) has a suite of workforce models that have been used in a variety of studies over many years to answer questions, generate forecasts and analyse various scenarios under consideration. Over time, these simulation models have been refined, errors corrected, and results validated against historical data. Consequently, the level of confidence in these models is very high. As computer technology advances, the need arises periodically to update these models to take advantage of newer technology and to provide more advanced capabilities. DRDC is updating its workforce modelling and simulation technology through the development of a Python-based discrete event simulation environment that is intended to replace various commercial simulation software products in which existing DRDC workforce models have been built. DRDC has begun the process of rebuilding select workforce models in this new environment. Rebuilding a simulation model using new technology usually results in some loss of confidence in the model even if the features of the new technology are impressive. This is due to the possibility of reintroducing errors, inexperience with the new technology, and a lack of validation against hard data. To establish confidence in the new model implementation, it is necessary to validate its equivalence to the older trusted version. This can be done by subjecting both implementations to identical input scenarios and comparing simulation output. However, because workforce models frequently make use of random effects modelled using probability distributions (e.g. age at recruitment, release events, and course failures), variability in simulation output from the same input is expected. Therefore, the challenge is to determine whether this variability is A) due to the known random effects that are part of the simulation model, or B) an indication that the behaviour of the new model implementation is not equivalent to the old. In this paper, we demonstrate the use of non-parametric and time series statistical techniques to test the hypothesis that the results obtained from Arena-based and Python-based implementations of the same military rank structure model are equivalent. Important measures of system performance that were tested include population size, time in rank, promotions, releases and course qualifications. This methodology will be of broad use to analytics practitioners who face the challenge of testing whether two independently-developed simulation models of the same system are in fact statistically equivalent.

**Keywords:** workforce model; simulation; Python; non-parametric statistics; time series; statistical validation

---

## 1. Introduction

### 1.1 Background

The management of military human resources (HR) in the Canadian Department of National Defence (DND) is a complex area in which decision-making affecting the careers of military personnel often produces unforeseen and long-lasting consequences. Without careful analysis, a decision to change personnel force structure, training programs, or operational commitments can unknowingly produce shortages and/or excesses of personnel in various occupations at various times and ranks and qualification levels. Ultimately, these unintended consequences may result in a reduction of or interruption in operational readiness.

In order to support the decision-making process in this domain, Defence Research and Development Canada (DRDC) develops models of military HR systems and conducts simulations in order to generate forecasts of the system effects of potential courses of action. Examples of HR-related information produced by this work include forecasts of: training demand, training output and occupation health (Corbett 2013; Latchman and Hunter 2002; Straver, Okazawa, and Wind 2009; Séguin 2011); manning levels at all ranks in various occupations (Zegers and Isbrandt 2010); and readiness for deployed operations (Scales, Okazawa, and Ormrod 2011; Moorhead, Wind and Halbrohr 2008).

Military HR planning occurs on an ongoing basis and is very dynamic due to changing policies, priorities, budgets, and political and economic events. Operating in this context, military HR models are often used

repeatedly and modified frequently over a long lifespan. Over this lifespan, simulation objectives typically evolve to become more ambitious, and computer hardware and software capabilities consistently advance at a rapid pace. Therefore, it is inevitable that older model implementations must eventually undergo significant upgrades to exploit technological developments in order to continue to meet the evolving requirements of simulation exercises.

DRDC is currently in the process of updating its workforce modelling and simulation technology through the in-house development of a novel discrete event simulation (DES) environment called the Right Person, Right Qualifications, Right Place, Right Time, Human Resources (R4 HR) simulation environment. This Python-based application is intended to replace various commercial simulation software products in which existing DRDC workforce models have been built. Several such models were originally built in the Arena DES software. While commercial applications like Arena are powerful, they generally cater to a wide audience in industry and are therefore not specialized for the particular needs of military HR simulation. Some of these specialized needs are the ability to easily code custom model logic, the ability to use relational databases and database querying during simulation execution, and the ability to interconnect separately developed models in order to conduct integrated simulation exercises (Okazawa 2013). These are the gaps that are addressed by the R4 HR simulation software.

DRDC has begun the process of rebuilding select workforce models in this new environment. The models chosen to be upgraded have an established record of informing military decisions over many years. The new versions of these models benefit from the specialized features of the R4 HR platform which facilitate the modelling process and provide more room for improvement as simulation requirements evolve. However, before the new versions can be used to make recommendations in real decision-making scenarios, they must undergo a validation process. The validation must demonstrate empirically that, subject to identical input scenarios, the new model implementation produces equivalent results to the older version in which confidence has already been established.

## **1.2 Aim**

The problem of demonstrating the equivalence of two different implementations of the same stochastic simulation model is the subject of this paper. As an example, we use a DRDC military rank structure model that was previously implemented in Arena and saw extensive use over many years, informing real-world military decision making in DND. This simulation model has now been re-implemented in the R4 HR software. The two model implementations, while logically equivalent, are not direct “line-by-line” translations of each other: they use different data structures, different programming techniques, and different random number generators for simulating stochastic processes. As a result of the latter, variation in simulation output is expected. However, it is necessary to determine whether these variations are just different random outcomes of the same underlying stochastic processes, or an indication of non-equivalence of the two model implementations. In order to test the hypothesis of equivalence, non-parametric and time series statistical techniques are used. Failure to reject this hypothesis provides validation of the equivalence of the two model implementations.

The paper is organized as follows. First, we briefly introduce the R4 HR simulation environment, identifying the features it provides that are suited to conducting military HR simulations, and indicating how the underlying concepts fundamentally differ from those of conventional DES applications. We then describe the military rank structure model, and highlight the differences in implementation between the Arena and R4 HR versions of the model that result from the different DES approaches. The methodology for statistically testing equivalence of the model implementations is then presented. Finally we discuss the test results which indicate, to a high degree of confidence, that the outputs of the two model implementations are statistically equivalent.

## **2. Python-based simulation environment**

The military HR models developed by DRDC to inform DND’s decision-making processes have typically been built in general purpose simulation applications, particularly Arena. While powerful, certain aspects of these applications were found to be not well-suited to the unique modelling requirements of military HR problems. The objective of the R4 HR software is to address these issues in order to make it easier to build military HR models and to reuse and modify them over their lifespan. Specifically, the main goals of R4 HR software are to provide:



- the flexibility of a programming language for defining model logic but without the steep learning curve normally associated with writing custom model code;
- the scalability of a relational database for storing simulation data and the ability to query the data using structured query language (SQL) during simulation execution; and
- a modular architecture such that model logic is easily reusable and interchangeable and allows individual models to be interconnected to create larger, integrated models.

As of the date of this publication, DRDC has developed a functioning prototype of the R4 HR simulation environment, and is investing in the development of a commercial grade version for use within DRDC and distribution to sister organizations in allied nations. This new software provides a number of unique features designed to meet the goals outlined above.

In R4 HR, simulation events are defined by writing compact routines in Python. These routines, called “code parts”, are linked to each other and to other data elements to build the model. R4 HR does not provide the standard set of DES building blocks such as assign and decide blocks; instead, it allows the user to write equivalent Python code. The rationale for this is that, in practice, the user-defined instructions that must be entered into conventional DES building blocks quickly become more complicated and difficult to manage than plain code. Thus, we opt for implementing model logic in code from the start. This requires only a basic level of Python programming knowledge for simple models, but also allows the logic to seamlessly scale to increasingly complex behaviour by adding to the code. Python was chosen because it is a general purpose, interpreted programming language that is used extensively in scientific research. It is powerful enough for large-scale scientific and web applications, while also having a straightforward syntax that is accessible to non-programmers. Furthermore, it is supported by an impressive variety of scientific libraries including statistical modeling, machine learning, large-scale data processing, and scientific plotting.

R4 HR includes an integrated relational database allowing simulation data to be stored in database tables and queried using SQL during simulation execution. For military HR models, the ability to store and process large volumes of data is essential. In many cases, relational databases and SQL are the ideal tools to handle personnel data during simulations.

R4 HR provides a modular model architecture through the implementation of a graph-alias naming system (GANS) (Okazawa 2013). In GANS, interactions between the elements of a model (i.e. the code, data, and other simulation objects) are handled explicitly using one-to-one connections rather than implicitly using scopes. Scopes are used in existing simulation software to define the set of model elements that can be accessed by a given element of model logic or line of code. For example, in Arena, data arrays, queues and the simulation clock are part of the global scope and are therefore accessible anywhere in the model. The disadvantage of scopes is that all names in a scope must be unique. If more than one element uses the same name, the name becomes ambiguous (known as a naming collision). Thus, it is generally not safe to select a group of model elements and duplicate it within the same model or move it into a new model because of the likelihood of creating naming collisions. In GANS, any model logic (up to and including the entire model) can be duplicated as-is and reused and reconnected with any other model because the connections between model elements are explicitly defined and are preserved with the model logic. Thus, duplicate names, if they occur, do not produce a naming collision.

R4 HR also facilitates modularity and model reuse by eliminating the concept of a global simulation clock and instead allowing the model developer to create multiple local clocks. The developer can independently set each clock’s time and tick rate. As the simulation proceeds, all clocks advance together in accordance with their relative tick rates. When a model element makes use of time (e.g. to create a delay), it must be connected to a local clock, and it then specifies points in time as measured by that clock. In this way, when a developer builds a large model by interconnecting several sub-models, each sub-model brings its own clock which may operate in different ways. For example, one sub-model may use a clock that ticks once a year and where the zero-time is the year 2000, while another sub-model may use a clock as a timer that ticks once a day and where the time is periodically reset to zero by the model logic.

Finally, R4 HR generalizes the concept of the conventional DES entity. Conventionally, an entity contains information that describes one instance of an object moving through a process (e.g. a parcel moving through the mail system), and entities are the primary information carriers in the model. However, in larger models,

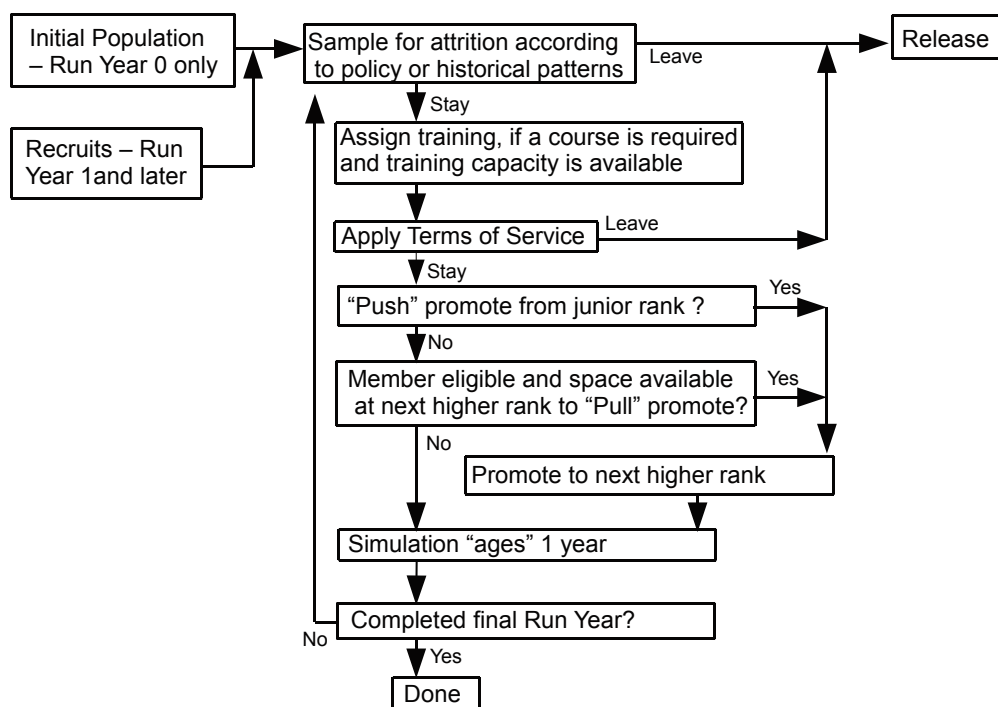
there are often many types of information that must be communicated, some of which do not fit the entity concept well. For example, one sub-model in a military HR model may request information from another sub-model regarding the availability and qualifications of a list of military personnel. This type of communication does not readily fit the entity concept. To address this, R4 HR provides a single mechanism that allows any information to be transmitted within the model as needed including simple values (numbers and strings), compound types (lists and dictionaries), custom objects, and traditional DES entities.

These fundamental features of R4 HR allow model developers to easily build models that are flexible, scalable and modular without requiring advanced knowledge in programming or modelling techniques.

### 3. Military rank structure model

The military rank structure model focusses on career progression in a military occupation, or group of occupations. The model is rank-oriented and rank-driven, and simulates members in a military occupation for up to 20 years forward. The projected number of promotions and the number of people leaving various ranks typically provide the drivers from which other aspects follow, such as the expected number of recruits needed, and the projected demand for training courses.

A schematic of the simulation model logic is provided in Figure 1. As a simulation proceeds from year to year, there is a sequence of events that changes the status of members in an occupation according to rules which are defined by career progression policies and the occupation's organizational structure. Once each run year, attrition rates are sampled from user defined probability distributions, and applied to the current population to determine releases at each rank level. After attrition has been applied, any required training is assigned to the remaining members of the population, taking into account factors such as prerequisite courses, course priorities, and course capacities. Training is applied at the lowest rank first, and then proceeds up through the ranks. The next event is the application of terms of service conditions; at pre-defined career points, members are offered a new service contract or are released. The next step is to apply promotions. Promotions into the senior ranks are based upon time served and qualifications held, and are subject to there being a vacancy at the next rank. Promotion into the junior ranks is based upon time served and qualifications held; no vacancy at the next rank is required. New recruits are introduced at the lowest rank level. Simulation information and outputs are recorded at various points during, and at the end of, each model run year.



**Figure 1:** Simulation logic for the military rank structure model

From the recorded simulation output, various measures of occupation health are tallied by model run year. The projection of occupation health over time provides expected trends in a number of areas (e.g. promotion

and release rates over time, average time in rank over the years, etc.) which can assist career managers with assessment of alternative management policies, as well as providing leading indications of potential future problems with overall occupational flows.

#### **4. Differences in model implementations**

The same model logic sequencing shown in Figure 1 is used in both the R4 HR and Arena implementations of the military rank structure model. The two model implementations employ the same assumptions (e.g. identical probability distributions for random events), and the mechanisms for the various simulation processes (such as recruitment, training and promotion) are logically equivalent. However, as the two modelling environments differ greatly, the R4 HR version cannot be a “line by line” translation of the Arena version. Some significant examples of the differences are as follows:

- In the Arena version, the model logic is largely implemented using standard DES building blocks. However, certain complex logic makes use of Arena’s embedded Visual Basic for Applications (VBA) coding to invoke custom behaviour. While the R4 HR version retains the same high level structure and sequence of events shown in Figure 1, it makes use of the Python code part and SQL database querying to implement the bulk of the lower level model logic.
- In the Arena version, members are represented by distinct entities which flow through defined paths in the model representing their advancement through the ranks and eventual release. In the R4 HR version, members being modelled correspond to rows in a database table, with their attributes stored in the table’s fields. This allows different portions of the model to easily access a member, and to apply and track changes (e.g. setting a new rank due to promotion).
- In the Arena version, while entities are used to carry information about individual military members, Arena’s signals (event triggers that broadcast to the whole model) are required to coordinate the flow of individuals moving through the occupation. For example, signals are used to ensure that all releases are processed before initiating training. In the R4 HR version, information about the status of individuals, and information used to coordinate their movement through the occupation are both transmitted by a single means: R4 HR’s generalized information transmission mechanism.

In addition to the above points, the two model implementations use different random number streams, so the same simulation processes will result in different specific outcomes for each run year, within each replication.

#### **5. Statistical validation methodology**

To validate the equivalence of the R4 HR and Arena implementations of the rank structure model, each version was used to analyse the same scenario, and the simulation outputs were compared. Given the stochastic nature of many aspects of the model (e.g. age at recruitment, release events and course failures), variability in the simulation output from the same input is expected. In order to determine whether the observed variability is A) due to different random outcomes of the same underlying stochastic processes, or B) an indication that the behaviours of the two model implementations are not equivalent, statistical non-parametric and time series techniques were used to compare the simulation outputs.

##### **5.1 Test scenario**

The military scenario utilized is fictional, and focusses on the management of a Maritime Operations occupation during a period of downsizing and capability transition. A demographic profile, incorporating attributes such as age, rank, time in rank, years of service, and qualifications held for each person in the starting population was created along with assumptions governing personnel policies, operational requirements and training capacity. The same input data and assumptions, which included both deterministic and stochastic elements, were used for the Arena and R4 HR implementations of the model.

Each implementation simulated 20 years of promotions, releases, recruitment, and training within the Maritime Operations occupation, and consisted of 50 model replications. Several personnel system performance measures were recorded, of which five that are key determinants of military occupation health were selected for the validation analysis:

- Population strength (the size of the Maritime Operations occupation at each rank level);
- Average time in rank for each rank level;
- Promotions to each rank level;

- Releases at each rank level; and
- Qualification strength (the number of people with particular qualifications at each rank level).

## **5.2 Non-parametric statistical tests**

For each of the performance measures selected, the simulation output consisted of two data sets (one from each model implementation), each containing 50 replications of a 20-year simulation period. As expected, the data sets generated did not follow normal distributions, negating the use of many parametric two-sample statistical tests to compare the data samples. To avoid making distribution assumptions that may not be valid, non-parametric techniques were employed to test equivalence of the outputs from the two model implementations. While very conservative, the Kolmogorov-Smirnov test was selected, as the only assumptions required are that the two samples are mutually independent, and come from continuous populations (Hollander and Wolfe 1973), assumptions which are satisfied. More powerful tests, such as the Wilcoxon rank sum test for location and Ansari-Bradley test for dispersion, require assumptions regarding the dispersion and location, respectively, of the continuous populations (Hollander and Wolfe 1973), assumptions that may or may not be valid. The null hypothesis for the Kolmogorov-Smirnov test is that the two samples come from the same continuous population, with the alternative hypothesis being they do not.

The Kolmogorov-Smirnov testing was conducted on each performance measure independently. For a given performance measure, the test was conducted 20 times—once for each model run year—with each test comparing the sample of 50 observations generated by R4 HR with the corresponding sample of 50 observations from Arena. Given that multiple hypothesis tests were being conducted, the Holm-Bonferroni method (Holm 1979) was used to control the family-wise Type I (false positive) error rate at level  $\alpha = 0.05$ .

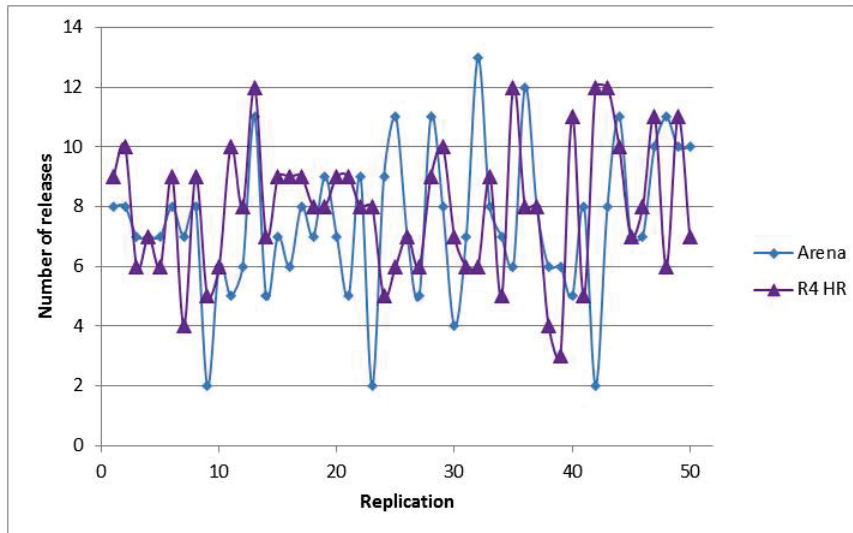
## **5.3 Time series analysis**

The non-parametric tests ignored the fact that the simulation output data were time series. For each performance measure, there were 50 independently generated time series covering a 20-year time period from each of the two model implementations. As the focus of the analysis was on pattern comparison rather than model fitting, formal statistical tests which require the fitting of time series models (e.g. autoregressive and/or moving average) to the data were not conducted.

Assuming the two model implementations are equivalent, the time series generated by the two implementations would follow the same patterns. This includes reproducing the same pattern of autocorrelations between model run years for a given performance measure. For each individual time series generated by the simulations, the sample autocorrelation coefficients at lag  $k$ ,  $k = 1, \dots, 20$ , were calculated. Some of the observed time series contained trends, however these were not removed prior to calculating the sample autocorrelation coefficients, as the intent was to compare existing patterns rather than fitting time series models to the data in order to predict future patterns. For each set of 50 time series, the 5<sup>th</sup>, 50<sup>th</sup>, and 95<sup>th</sup> percentiles of the sample autocorrelation coefficients were calculated, and plotted. Any visually notable deviations between the percentile patterns for corresponding data samples were interpreted as indications of non-equivalence of the two model implementations.

## **6. Validation Results**

As noted above, many performance measures were analysed for the validation exercise. One representative measure is the number of releases annually at the rank of Lieutenant (Navy) (Lt(N)) from the Maritime Operations occupation. The rank of Lt(N) can be a career progression plateau for some members, resulting in population sizes appropriate for the conduct of statistical tests. Figure 2 shows the number of Lt(N) releases observed across all replications for a single model run year. As expected, for any given run year there is variability in the results produced by the two model implementations.



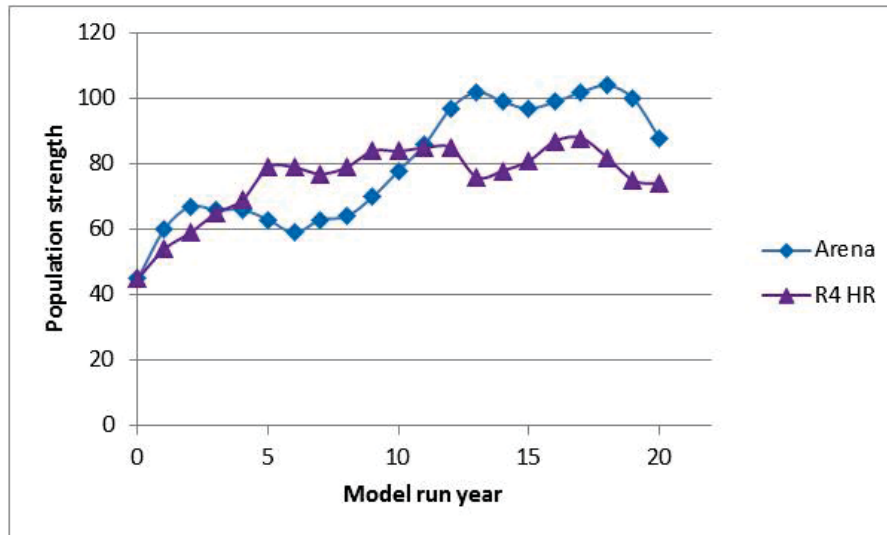
**Figure 2:** Number of Lt(N) releases observed in model run year 10

Applying the Kolmogorov-Smirnov two-sided test to the data samples of Lt(N) releases per run year generated by the two model implementations yields the results shown in Table 1. For model run year 10, the observed variability pictured above is not significant enough for the Kolmogorov-Smirnov test to reject the null hypothesis of a common population distribution for the two samples. The smallest p-value observed across the 20 multiple tests is 0.0171 in run year 15. As this value is greater than  $0.05/20$ , the Holm-Bonferroni method does not reject any of the hypotheses. The same result was observed for all performance measures examined—in all cases the null hypothesis of a common population distribution for the Arena and R4 HR generated data samples could not be rejected. Failure to reject the hypotheses of common population distributions supports the conclusion of statistical equivalence of the two implementations of the military rank structure model.

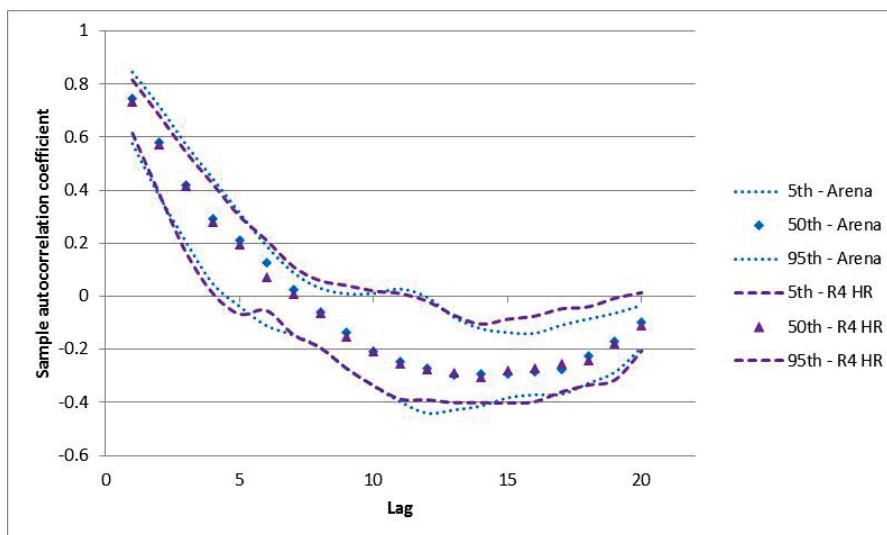
**Table 1:** Kolmogorov-Smirnov two-sided test results for Lt(N) releases by model run year

Run year	Test statistic	p-value	Run year	Test statistic	p-value
1	0.08	0.9958	11	0.18	0.3584
2	0.0531	1	12	0.1	0.9541
3	0.08	0.9958	13	0.1	0.9541
4	0.12	0.8409	14	0.18	0.3584
5	0.1	0.9541	15	0.3	0.0171
6	0.08	0.9958	16	0.14	0.6779
7	0.14	0.6779	17	0.2	0.2408
8	0.08	0.9958	18	0.14	0.6779
9	0.16	0.5077	19	0.16	0.5077
10	0.16	0.5077	20	0.12	0.8409

The second component of the validation exercise took into account the time series nature of the data. Figure 3 shows the number of Lt(N) in the Maritime Operations occupation over time as generated by the two model implementations for a single replication. In this case the population is growing as the number of individuals entering the rank annually exceeds the number leaving. The 5<sup>th</sup>, 50<sup>th</sup> and 95<sup>th</sup> percentiles of the observed autocorrelation coefficients for the two model implementations are shown in Figure 4. No visually noteworthy differences in the autocorrelation patterns between the two simulation output data sets are discernable. Similar results were observed for all performance measures analysed. These results provide further evidence to conclude that the R4 HR and Arena implementations of the military rank structure model are equivalent.



**Figure 3:** Number of  $Lt(N)$  in the Maritime Operations occupation over time for replication 5



**Figure 4:** Select percentiles of the sample autocorrelation coefficients at lag  $k$  for  $Lt(N)$  population strength

## 7. Conclusion

The development of the R4 HR simulation environment is enabling DRDC to update its workforce modelling and simulation technology. This paper provided brief overviews of the R4 HR application and a well-established and trusted military rank structure model that has been rebuilt in the new environment. In order to establish confidence in the new model implementation, a validation process was undertaken. As described in this paper, the old (Arena) and new (R4 HR) model implementations were used to analyse the same fictional military HR scenario. Non-parametric statistical tests and time series analysis techniques were used to test statistical equivalence of the simulation outputs, and by association, the logical equivalence of the two model implementations. The results of Kolmogorov-Smirnov tests and examinations of autocorrelation patterns for each of several measures of system performance did not detect any statistically significant differences in the simulation outputs. We thus conclude that the R4 HR model implementation is indeed logically equivalent to that of the Arena model, which will assist with building user and client confidence in the new R4 HR simulation environment. The approaches presented here can be deployed readily by analytics practitioners facing the challenge of statistically validating the equivalence of independently-developed simulation models of the same system.

## References

- Corbett, N. (2013) *Modelling the Production and Absorption of Pilots: The Development of the Production, Absorption and Retention Simulation (PARSim)*. DRDC Technical Report 2013-023, Centre for Operational Research and Analysis, Ottawa, Canada.
- Hollander, M. and Wolfe, D. (1973) *Nonparametric Statistical Methods*. John Wiley & Sons, New York.
- Holm, S. (1979) "A simple sequentially rejective multiple test procedure", *Scandinavian Journal of Statistics*, Vol 6, No. 2, pp. 65 – 70.
- Latchman, S. and Hunter, C. (2002) *Preliminary Results from the Pilot Production/Absorption/Retention Simulation (PARSim) Model*. 1 CAD/CANR Headquarters Research Note 0202, Centre for Operational Research and Analysis, Ottawa, Canada.
- Moorhead, P., Wind, A. and Halbrohr, M. (2008) "A Discrete Event Simulation Model for Examining Future Sustainability of Canadian Forces Operations", *Proceedings of the 2008 Winter Simulation Conference*, S.J. Mason, R.R. Hill, L. Mönch, O.Rose, T. Jefferson, and J.W. Fowler (ed.), pp. 1164 – 1172. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
- Okazawa, S. (2013) "A Discrete Event Simulation Environment Tailored to the Needs of Military Human Resources Management", *Proceedings of the 2013 Winter Simulation Conference*, R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M.E. Kuhl (ed.), pp. 2784 – 2795. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
- Scales, C., Okazawa, S. and Ormrod, M. (2011) "The Managed Readiness Simulator: A Force Readiness Model", *Proceedings of the 2011 Winter Simulation Conference*, S. Jain, R.R. Creasey, J. Himmelsbach, K.P. White, and M. Fu (ed.), pp. 2519 – 2529. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
- Seguin, R. (2011) "1 Canadian Forces Flying Training School (1 CFFTS) Resource Allocation Simulation Tool", *Proceedings of the 2011 Winter Simulation Conference*, S. Jain, R.R. Creasey, J. Himmelsbach, K.P. White, and M. Fu (ed.), pp. 2495 – 2506. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
- Straver, M., Okazawa, S. and Wind, A. (2009) *Training Pipeline Modelling Using the Production Management Tool*. DRDC Technical Memorandum 2009-019, Director General Military Personnel Research and Analysis, Ottawa, Canada.
- Zegers, A. and Isbrandt, S. (2010) "The Arena Career Modelling Environment – A New Workforce Modelling Tool for the Canadian Forces", *Proceedings of the 2010 Summer Computer Simulation Conference*, G. Wainer (ed.), pp. 94 – 101. Curran Associates, Inc.



# Development of a Naval Reserve Training and Career Progression Analysis Application

**Michelle Straver**

**Director General Military Personnel Research and Analysis, Ottawa, Canada**

[michelle.straver@forces.gc.ca](mailto:michelle.straver@forces.gc.ca)

**Abstract:** Within the Canadian Armed Forces (CAF) Naval Reserve (NAVRES), attrition has outpaced recruitment since 2010. In the current fiscal environment, resources for recruiting are limited; this further increases the pressure to improve retention to ensure a sufficiently large workforce. For some years now, there has been concern that personnel tend to leave the organization when they are unable to commit the time required to complete the training necessary for career advancement. This concern is particularly applicable to part-time personnel, who are more likely to have demanding personal or professional commitments outside of the NAVRES (e.g. a full-time civilian job). A stronger understanding of members' career progression patterns and challenges, such as major barriers in completing training, would help the NAVRES to identify key target areas for improving the efficiency of the system such that personnel would be more likely to remain in the NAVRES and advance further in their careers. A user-friendly, flexible application was developed to enable NAVRES planners to investigate personnel management questions. The application enables the user to conduct a cohort analysis to trace the career path of each person in a population of interest, from the time of joining the population to the time of leaving (or up to the time that the analysis is carried out). In doing so, the user is able to address questions such as: What percentage of personnel achieves a particular qualification? How long does it take for personnel to achieve that qualification? After a certain number of months or years of service, how many of the key training courses have personnel completed? Among those who left the population of interest within a certain period of time, how many remained in a different element of the CAF, and how many left the CAF entirely? Further, the answers to these questions can be compared across different subgroups of the overall population; e.g. for full-time vs. part-time personnel, or across personnel serving in different occupations. This paper will provide an overview of the application, and will address some of the challenges that were encountered, such as dealing with incomplete or inconsistent data. An example focusing on a specific occupation will also be provided as a demonstration of the breadth of results that can be obtained using the tool.

**Keywords:** workforce modelling; cohort analysis; training; career progression

---

## 1. Introduction

### 1.1 Rationale

Within the Canadian Armed Forces (CAF) Naval Reserve (NAVRES), attrition has outpaced recruitment since 2010. In the current fiscal environment, resources for recruiting are limited; this further increases the pressure to improve retention to ensure a sufficiently large workforce. In recent years, there has been concern within the NAVRES that personnel tend to leave the organization when they are unable to commit the time required to complete the training necessary for career advancement.

This concern is particularly applicable to part-time personnel, who are more likely to have demanding personal or professional commitments outside of the NAVRES (e.g. a full-time civilian job). Indeed, several previous studies have found that part-time reservists have limited availability to undergo full-time training (Jenkins 2005a; Mehmood 2007); and further, that outside commitments discourage some personnel from pursuing full-time service (Jenkins 2005b). A stronger understanding of members' career progression patterns and challenges, such as major barriers in the training system, would help the NAVRES to identify key target areas for improving the efficiency of the system such that personnel may be more likely to remain in the NAVRES and advance further in their careers.

### 1.2 Previous Research

Unlike most civilian organizations, in military organizations, personnel generally join at the entry level having little or no experience in the military domain, and progress to higher ranks as they gain knowledge and

experience (Boileau 2012). A consequence of this type of system is that decisions (e.g. changes to training systems, policy changes, etc.) can have long-term consequences. Thus, to ensure an adequate supply of trained personnel, military organizations must conduct workforce planning activities on an ongoing basis (Wang 2005), whether through data analysis, modelling and simulation, or other techniques. A large volume of research has been done in this area, covering a wide variety of techniques and subjects. While only very little published research has pertained to the NAVRES in particular (e.g. the systems dynamics model developed by Mehmood (2007)), if we look beyond the NAVRES, we find that some previous work has elements in common with this study.

Training is an important element of a military member's career development, so workforce planning activities in this area are of great interest to decisionmakers. Within the Canadian military context, workforce planning research pertaining to training often examines the early stages of one's career (Christopher 2009; Latchman and Michaud 2010; Straver et al. 2009), but can also focus on training at more advanced levels (Moorhead 2006; Straver 2012a; Woodill 2005).

Focusing on analysis techniques, while cohort analyses are very common in the medical field, this approach also has workforce planning applications. Within the CAF, it has been used to examine issues such as calculating early-career attrition (Latchman and Michaud 2010), measuring the time to completion of basic occupation training (Straver 2012b), and estimating the cost of implementing a change to promotion policy (Straver, Latchman, and Tabbenor 2013).

### **1.3 Aim**

The aim of this study was to enable NAVRES planners to better understand training and career progression patterns and challenges within each NAVRES occupation. This was done by developing an application that would enable the user to obtain the answers to a variety of questions, including (but not limited to):

- What percentage of personnel achieves a particular qualification?
- How long does it take for personnel to achieve that qualification, and has this changed over time?
- After a certain number of months of service (MOS), how many of the key training courses have personnel completed?
- Among those who left the population of interest within a certain number of MOS, how many transferred to a different occupation in the NAVRES, how many transferred to the Regular Force or another element of the CAF, and how many left the CAF?

Further, the application would allow the user to compare the answers to these questions across different subgroups of the overall population; e.g. for full-time vs. part-time personnel, or across personnel serving in different occupations.

## **2. Methodology and Application Development**

This section outlines the overall approach used for the analysis, as well as details on the development of the application.

### **2.1 Cohort Analysis**

To answer the questions stated previously, it is necessary to trace individuals through their career from the point of joining the occupation to the point of leaving (or to the date of the most recent available data set). A cohort analysis was identified as the best way of solving the problem.

In a previous study (Straver, Latchman, and Tabbenor, 2013), the CAF population hired within a specific time period was tracked through 72 months of service (MOS) and separated according to whether they remained in the population or not; those who remained were further separated by rank. Here, a similar approach was taken, though at a higher level of detail.

## 2.2 Building the Application

The decision was made to provide the NAVRES with an application that staff could use themselves. This approach has two main advantages over having DGMPRA conduct the analysis and provide the end results to NAVRES staff:

- NAVRES staff can use the application to answer a wider variety of questions, rather than having DGMPRA focus on answering only the highest-priority questions.
- With minimal support from DGMPRA, NAVRES staff will be able to re-use the application for years to come as newer data become available. This could be done to, for example, evaluate the impact of a change to the training system by comparing the results from before and after implementation.

These advantages came at only minor costs during the development of the application. First, additional time was required to implement rules-based solutions to account for data irregularities, rather than applying “band-aid” solutions (e.g. manual correction of errors). Additionally, it was necessary to take the time to design a user-friendly interface, and to provide supporting documentation. Importantly, the documentation needed not only to include instructions for use, but also to identify the application’s assumptions and limitations.

An additional constraint was that the application needed to be built in an environment that NAVRES staff would have access to, i.e. using commonly available software. Although some of the pre-processing of the data was done in an MS Access database, the majority of the analysis, along with the user interface itself, was implemented in MS Excel.

The remainder of this section will provide an overview of how the application was built, from a description of the necessary data to the user interface. Sample results will also be provided to demonstrate the key metrics that can be obtained.

### 2.2.1 Data Source

All data used for this study were taken from the Naval Reserve Information Management System (NRIMS), updated to 27 May 2013. This database contains information on each member of the NAVRES, going back to the 1990s. For this analysis, all of the required information was extracted from seven key tables:

- *MOS\_HIST*. This table contains data on the occupation(s) that each person served in (or is serving in), including start and end dates (if applicable) for each case.
- *ELEMENT\_HIST*. This table gives the time periods (i.e. start and end dates) during which each person served in an element of the CAF. In this study, the element of greatest interest is the NAVRES. However, personnel may have spent time in a different element before joining or after leaving the NAVRES, such as in another primary reserve element (e.g. the Air Reserve or the Militia), the Supplementary Reserve, or the Regular Force.
- *ENROLMENT\_RELEASE*. Although this table includes information on enrolments as its name suggests, only data on releases were used; most notably release dates, and to a lesser extent release items (i.e. release reasons, such as voluntary or medical). Of interest, a release date often (but not always) exists even in cases where an individual transferred from the NAVRES to a different element within the CAF.
- *EMP\_HIST*. This table was used to identify the time periods during which each member worked on a Class B or Class C contract. Reserve service falls into three classes. Class A service is considered part-time and is applicable when the member serves for periods of 12 consecutive days or less. Class B service is applicable when the individual is on full-time service of a temporary nature, for periods of 13 consecutive days or more. Class C service refers to full-time service that is not of a temporary nature (DSCS, 2006). Reservists will frequently move between the three classes. As will be discussed later, as an approximation for the purposes of this study, the time spent on Class B and Class C contracts is used to differentiate between part-time and full-time personnel.
- *QUAL\_HIST*, *QUAL\_VIEW*, and *CAREER\_PROG*. Personnel will complete a number of courses during their career, some of which will lead to the achievement of particular qualifications that are required for career progression. The *CAREER\_PROG* table specifies the normal sequence of qualifications applicable to each occupation; together with *QUAL\_HIST* and *QUAL\_VIEW*, the dates of achievement of these qualifications can be obtained for each person.

### 2.2.2 Pre-processing of Data

The first step of the analysis was to pre-process the raw data from the NRIMS tables listed above to extract the required data elements required for each person in the data set, and to clean the data in doing so. Each data element is described below.

- *Occupation.* Within the CAF, changes are frequently made to occupation structures; occupations may merge, divide, or be renamed. All obsolete occupations were translated to their current equivalents.
- *Start Date.* For each combination of service number (SN) and occupation, the first instance of the start date in the occupation was identified, regardless of which CAF element the member was serving in at the time (e.g. NAVRES, Regular Force, etc). Next, these data were filtered to include only the cases where the member was in the NAVRES at the time. This approach ensured that only personnel who were joining the occupation for the first time would be included in the analysis. Otherwise, a bias would be introduced by the inclusion of personnel who already had some qualifications in the occupation upon joining the NAVRES.
- *End Date.* For each instance of joining a NAVRES occupation as identified in the previous step, the time of leaving the occupation was determined. While this information was generally included in the MOS\_HIST table, there were a number of unexpected null values. For this reason, the end date was calculated as the minimum of the occupation end date (from MOS\_HIST), the element end date (from ELEMENT\_HIST), and the release date (from ENROLMENT\_RELEASE, filtered to capture the earliest instance that occurred after the date of joining the occupation). Some further validation was required. For example, cases were found where the MOS\_HIST table indicated that a person left an occupation on one day and restarted in the same occupation on the next day. In these cases, the data were modified such that a continuous period of service was reflected. Similarly, cases where an individual joined and left an occupation on the same day were excluded.
- *Destination.* Through manipulation of the MOS\_HIST, ELEMENT\_HIST, and ENROLMENT\_RELEASE tables, upon leaving an occupation, personnel were categorized according to whether they left the CAF and did not return, left the CAF but returned after a break in service, transferred to the Regular Force, transferred to another NAVRES occupation, or transferred to another CAF element. It is worth adding that information on the release type was also collected in case it would be required for future use; however, since the vast majority of releases were voluntary, this was not considered to be of great interest at this time.
- *Time on B/C Contract.* After obtaining the start and end dates of the time served in an occupation, the employment data from the EMP\_HIST table was used to determine how many of these days the person spent on a Class B or Class C contract, i.e. how many days were spent working on a full-time basis.

Once these data elements were obtained, they were filtered according to the occupation being analyzed, and exported into the application. Next, the qualifications associated with the occupation, along with the achievement dates, were exported for each person in the data set for further processing within the application.

### 2.2.3 Data Manipulation

The next steps of the analysis are done simultaneously from within the application as it is being used. Although this adds considerably to the file size, it has the advantage of transparency; this can help the user identify the reasons behind any counter-intuitive findings, and to correct for any remaining data anomalies that were not already removed in pre-processing.

After pre-processing, the data are structured as a table with each row representing a different person in the population. Each row contains the person's SN as well as the pertinent data elements. At this point, all data points are included in the data set, i.e. all personnel who joined the occupation for the first time since 1 January 1998. The data are then filtered according to the specifications set by the user (discussed later in this paper). An additional filter allows the user to exclude any rows that he/she does not wish to include; this is useful for excluding data anomalies.

Through the use of lookup tables to relate the qualification data to each person's dates of joining and leaving the occupation, the application calculates the MOS that each individual had at the time of achieving the qualifications required for career progression. In cases where the qualification was never achieved, or where

the qualification was achieved after the time of leaving the occupation, data are excluded by setting the MOS to an arbitrary high value (9999).

Based on these calculations, the highest qualification achieved as of each MOS point is determined for each person in the filtered data set. These are then summed to enable the overall results to be obtained.

A similar approach is taken to evaluate, at each MOS point, how many personnel remain in population, and to separate the “leavers” according to the destination groups described previously.

## 2.2.4 User Interface and Sample Results

In this section, the user interface is described, along with key metrics of interest to NAVRES staff. All examples are taken from the analysis of the Maritime Surface and Subsurface (MARS) occupation.

The user interface for the MARS occupation is shown in Figure 1. Labelled items are discussed below.

MARS (MOSID 00207)					Enter MOS point to be analyzed: 96			
Range for Occupation Start Date		1	01-Jan-98	31-Dec-04	(min = 1 Jan 1998, max = 27 May 2013)			
Range of Class B/C Days Worked / Days in Occ			0%	20%	Suggest 0%-20% to capture part-time personnel; 70%-100% to capture full-time personnel.			
Date of NRIMS Extract			27-May-13					
Latest Valid MOS Point			100	3				
2					4	5	6	7
Qualification	Order	BTL	Rank	Include in Graphs	Average Time to Completion (if achieved by 96 MOS) (Excludes those who had qual upon entry into occupation)	% Achieved Qualification (if achieved by 96 MOS) (Includes those who had qual upon entry into occupation)	% Achieved Qualification (if achieved by 96 MOS) (Includes those who had qual upon entry)	For Selected Qualifications: Completion Rate by 96 MOS
BMOQ	1	1	A-SLT	1	4.4	69%	100%	69%
NETPO DL	2	1	A-SLT	1	8.6	60%	100%	
NETPO F2F	3	1	A-SLT	1	8.3	58%	100%	84%
MARS 3	4	1	A-SLT	1	22.6	34%	100%	59%
MARS 4 PT1	5	1	A-SLT	1	29.2	32%	96%	
OOD DL	6	1	A-SLT	1	32.9	26%	100%	
MARS 4	7	1	A-SLT	1	37.5	22%	96%	63%
OOD AFLOAT	8	1	A-SLT	1	40.0	21%	96%	95%
BWIK	9	1	LT(N)	1	41.2	21%	96%	100%
BOC MOD 1	10	0	LT(N)	1	50.2	12%	58%	
BOC MOD 2	11	0	LT(N)	1	50.4	12%	58%	58%
MWBC	12	0	LCDR	1	61.0	8%	46%	
COPS MOD1	13	0	LCDR	1				
COPS MOD2	14	0	LCDR	1				
JCSF DL	15	0	CDR	1				
AEDP	16							
AEFG	17							

Figure 1: User Interface

In Item 1, the user specifies the range of occupation start dates that should be included in the cohort analysis; that is, personnel who joined the occupation outside of these dates are excluded. The dates may range from 1 January 1998 to the date of the NRIMS update. It is necessary for the user to select a wide enough range such that the data set is sufficiently large to be able to extract valid results; however, the user must also take into consideration the latest valid MOS point that will result. The latest valid MOS point is based on the difference between the upper end of the selected date range and the date of the NRIMS update; thus, it refers to the MOS served by those who joined the occupation on the last day of the selected range. Results can be shown only up to this MOS point. If, for example, the user selects a date range from 1 January 1998 to 31 December 2012, only data up to 4 MOS can be analyzed, since those hired on 31 December 2012 would only have 4 complete MOS as of 27 May 2013.

Also in Item 1, the user selects the range of the ratio of Class B or Class C days worked to the total number of days in the occupation. As noted previously, Class B and Class C reservists are considered full-time. As an approximation, a range of low ratios (e.g. 0% to 20%) will capture part-time personnel, whereas a range of high ratios (e.g. 70% to 100%) will capture full-time personnel.

Item 2 shows the sequence of qualifications in the career progression for the occupation, along with the rank associated with each qualification. The Basic Training List (BTL) column indicates whether the qualification is part of basic occupation training or more advanced training that would taken after an individual is considered qualified in his/her occupation. All of these are specified in the NRIMS; note that the application can accommodate up to 40 different qualifications. The user can then select which of these qualifications he/she wishes to be shown in the results. It may not be desirable to select all qualifications; rather, the user may wish

to focus on the qualifications that involve lengthy training courses, or to ignore the initial phases of training, or to focus on a specific rank, etc.

In Item 3, the user specifies the MOS point that is to be analyzed. This can be any MOS point up to the latest valid MOS point identified in Item 1.

Items 4 to 7 give some of the key metrics. Of note, all of these metrics pertain to qualifications achieved by the MOS point being analyzed, as specified by the user.

Item 4 gives the average time to completion of each qualification, which is based only on the personnel who did not have the qualification upon entry into the occupation. As noted above, no personnel in the data set should have any of the qualifications that are exclusive to the occupation being analyzed; however, some may have qualifications that are shared with other occupations, such as the Basic Military Officer Qualification (BMOQ).

Item 5 gives the percentage of the starting population that achieved each qualification. Item 6 is similar, but gives the percentage of the *remaining* population (i.e. of those who did not leave the occupation before the specified MOS point) that achieved each qualification.

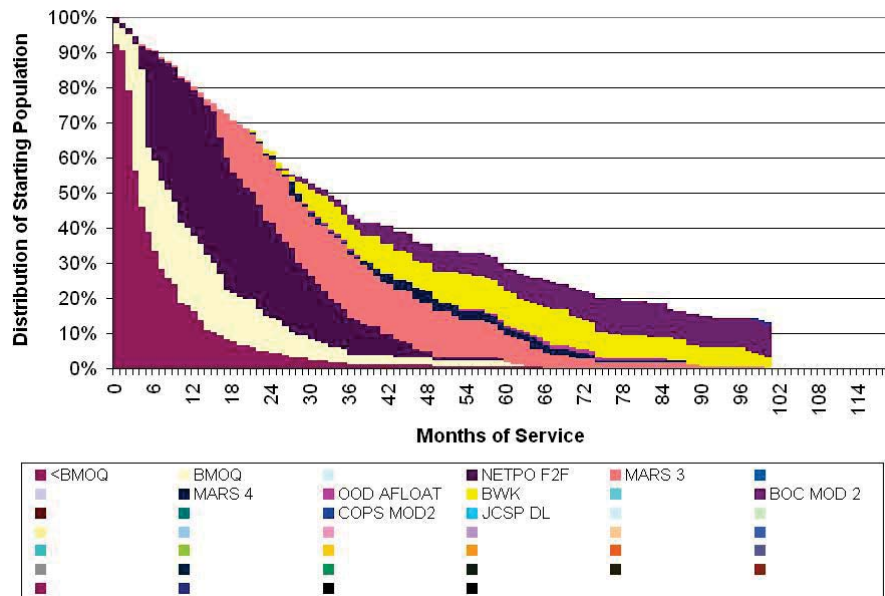
Item 7 gives what is perhaps the most interesting of this set of metrics: the percentage of those who achieved the qualification, *measured out of those who achieved the previous qualification*. This is particularly useful for identifying common “off ramps” in the training system. From Figure 1, we can see relatively low completion rates for MARS 3 (out of those who completed NETPO F2F); for MARS 4 (out of those who completed MARS 3); and BOC MOD 2 (out of those who achieved BWK).

Even more useful results can be obtained by comparing these results with those of a different population. For example, if we change the parameters to capture full-time personnel rather than part-time personnel (by changing the range of the Class B/C ratios in Item 1 from 0%-20% to 70%-100%), we obtain the following results: 92% of those who completed NETPO F2F also completed MARS 3 (vs. 59% of part-time personnel); 97% of those who completed MARS 3 also completed MARS 4 (vs. 63% of part-time personnel); and 94% of those who achieved the BWK also completed BOC MOD 2 (vs. 58% for part-time personnel). Results like this can help NAVRES planners determine which training courses are the most problematic for part-time personnel to complete.

Although the user can obtain the above metrics for any MOS point, certain metrics are better presented in graphical form. Two examples are given below.

For each MOS point up to the maximum valid value (100 in this example), Figure 2 shows the distribution of the starting population according to the highest qualification achieved. Only the qualifications selected by the user are included; blank series are a consequence of this flexibility.

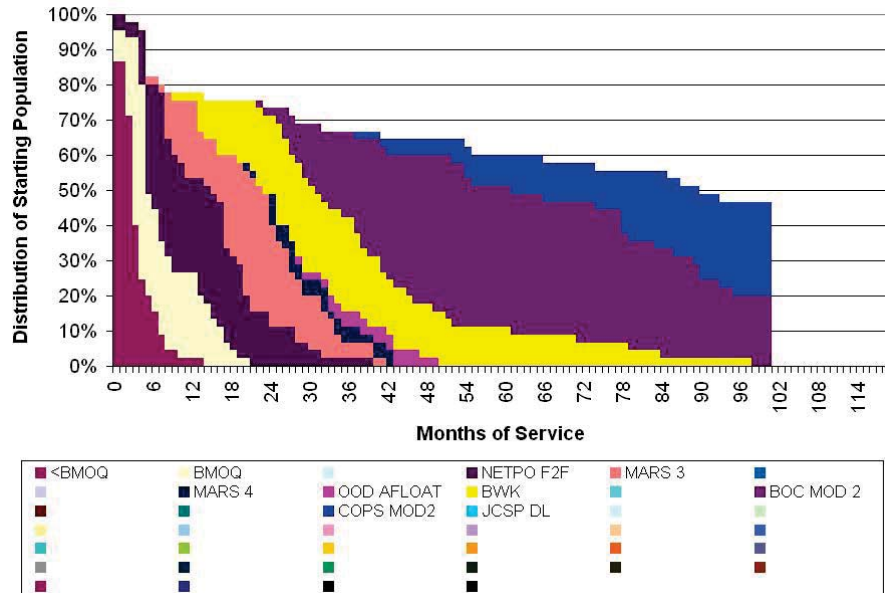




**Figure 2:** Highest Qualifications Achieved by MOS for Part-Time MARS Personnel

This display allows the user to draw several conclusions. First, since the distribution is based on the starting population, the sum of the series tells the user how many remain in the population at a given MOS point. The width of each coloured band can also provide insights on the length of time required to achieve a certain qualification (either overall, or in comparison to a previously-earned qualification).

Again, more interesting results can be obtained by comparing the results for different populations. Continuing from our previous example, Figure 3 shows the same graph for the full-time group of MARS officers.

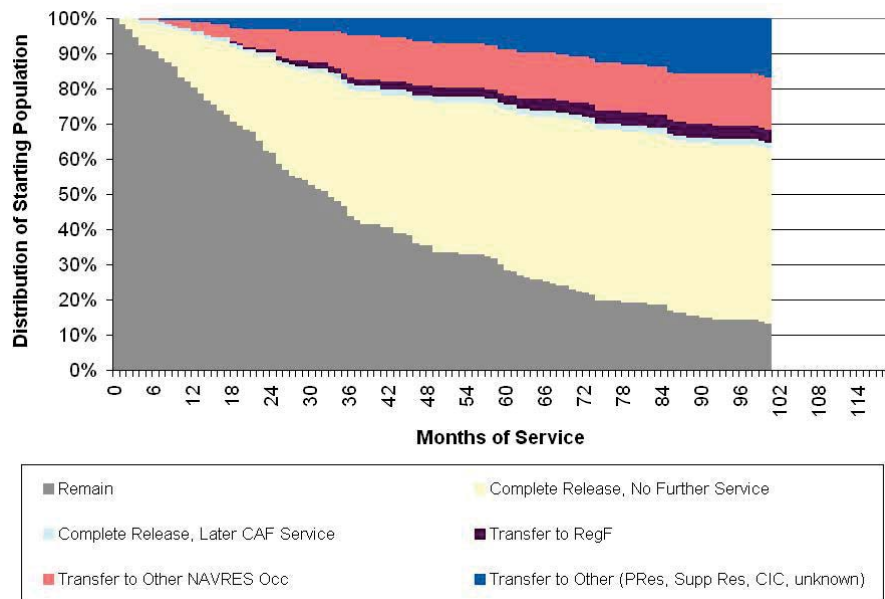


**Figure 3:** Highest Qualifications Achieved by MOS for Full-Time MARS Personnel

By comparing Figures 2 and 3, several conclusions can be drawn. For example, retention is generally higher for part-time personnel over the first 6 MOS, but lower thereafter. Also, not surprisingly, full-time personnel achieve qualifications in a shorter period of time than do part-time personnel; consequently, for a given MOS point, full-time personnel will tend to have more qualifications.

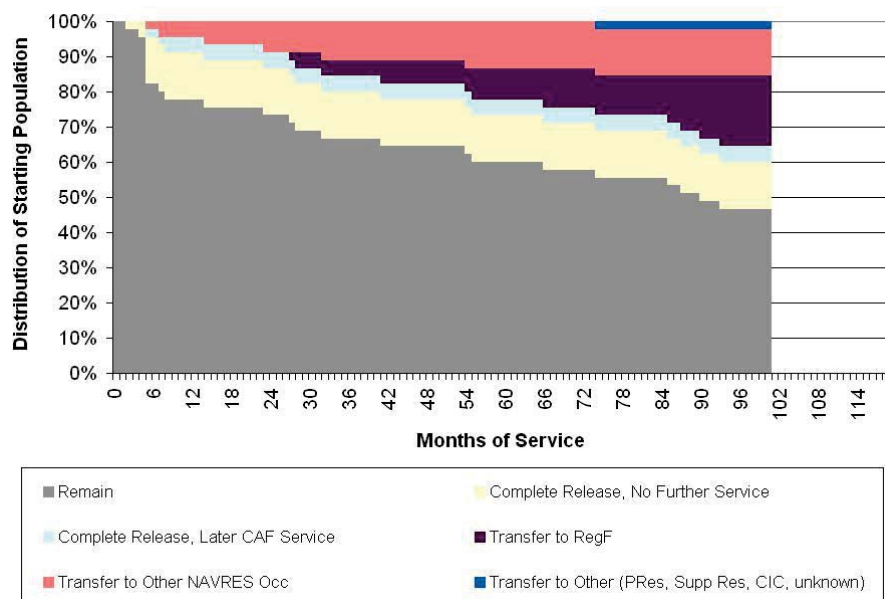


Figure 4 is opposite to Figure 2 in that the information is focused on the personnel who have left the occupation rather than the remaining personnel. It shows the distribution of these personnel according to the destination groups noted in Section 4.1.



**Figure 4:** Destination Groups by MOS for Part-Time MARS Personnel

Figure 5 provides similar information for the full-time population.



**Figure 5:** Destination Groups by MOS for Full-Time MARS Personnel

Continuing with our previous example, by comparing Figures 4 and 5, it can be observed that upon leaving the MARS occupation, full-time personnel are much more likely to transfer to the Regular Force, whereas part-time personnel are much more likely to transfer to another element of the CAF or to be released from the CAF altogether.

### 3. Notes of Interpretation

The key point that must be understood by the user when interpreting the results is that the data set only includes personnel who had no prior experience in the occupation. For this reason, results cannot be extended to the overall population. As an example, looking at Figure 3, the user might be inclined to conclude that, among full-time MARS personnel, approximately 18% have achieved BWK, 67% have achieved BOC MOD 2,

and 15% have achieved COPS MOD 2 by the 60 MOS point. However, this is not the case. Rather, it is likely that the overall population is more qualified as a result of personnel who enrol in the NAVRES with prior Regular Force or NAVRES experience (and by extension, qualifications) in the occupation. Similarly, the patterns observed in the destination groups that personnel go to upon leaving their NAVRES occupation could be different if personnel with prior service were to be included in the data set.

The user must also be careful when choosing the range of start dates that define the cohort to be analyzed. For example, a user may wish to assess the impact of a change to a course that was implemented in January 2010. It would be incorrect to define the pre-change ("before") cohort as the personnel who joined the occupation up to and including December 2009. This is because, although the selected personnel will have joined the occupation prior to the implementation of the change, they may not have taken the course until some time later. Instead, the user would need to select an earlier range of start dates, and examine the results to ensure that all (or at least most) of the personnel in the cohort had completed the course prior to the implementation date of the change.

#### **4. Conclusion**

This paper outlined the development of a user-friendly, flexible analysis application that is intended for use by NAVRES planners to investigate personnel management questions. In total, 15 versions of the application were built, each pertaining to a different NAVRES occupation.

Developing an application for NAVRES planners to use themselves had several advantages over the alternative of providing only end results: first, NAVRES staff will now have the capability of reusing the application as newer data become available; secondly, NAVRES staff will be able to answer unanticipated questions as they arise. These advantages came at only minor costs during the development process.

The overall process of building the application involved several steps to extract and pre-process the NRIMS data into a form that could be exported and used within the application. However, much of the analysis is done within the application itself, based on user inputs.

The example given in this paper demonstrated that NAVRES planners can use the application to answer questions on the subject of training and career progression; most notably, pertaining to the achievement of qualifications. In doing so, it can help planners to identify target areas for improving the efficiency of the training system. For example, it was found that only 59% of part-time personnel who achieved the NETPO F2F qualification also achieved MARS 3; in contrast, 92% of full-time personnel did so. This suggests that the MARS 3 course is problematic for part-time personnel, and would be a suitable place to start when seeking to increase the efficiency of the system.

The application is also useful for supporting or disproving certain hypotheses about training and career progression. Indeed, NAVRES planners were able to confirm that the average time to completion of basic occupation training for the MARS occupation was within their standard of 36 months (for part-time and full-time personnel together).

Finally, it is of interest to note that this work could have applications beyond the military context. Although the application is set up to suit the structure of the NRIMS database, the general approach could be adapted for use in other organizations. For example, it could help to identify common exit points and/or systemic barriers along a lengthy training system, such as an undergraduate university program, training systems for medical or legal professionals, etc.

#### **References**

- Boileau, M. (2012) "Workforce Modelling Tools Used by the Canadian Forces", *Proceedings of the International Workshop on Applied Modeling & Simulation*, Rome, Italy, September.
- Christopher, G. (2009) *Implications of ROTP Recruiting*, Briefing Note, Director General Military Personnel Research and Analysis, Ottawa, Canada.
- Directorate Strategic Corporate Services (DSCS). (2006) *Queen's Regulations and Orders (QR&O) for the Canadian Forces*, Vol 1, Articles 9.06-9.08.
- Jenkins, D.A. (2005) *Naval Reserve Recruiting and Retention Project: Phase One*, Sponsor Research Report 2005-01, Director Human Resources Research and Evaluation, Ottawa, Canada.

- Jenkins, D.A. (2005) *Naval Reserve Recruiting and Retention Project: Phase Two*, Sponsor Research Report 2005-15, Director Human Resources Research and Evaluation, Ottawa, Canada.
- Latchman, S. and Michaud, K. (2010) "Early Attrition of Army Non-Commissioned Members (NCMs)", Presentation given at 52<sup>nd</sup> Annual International Military Testing Association Conference, Lucerne, Switzerland, September.
- Mehmood, A. (2007) "Application of System Dynamics to Human Resource Management of Canadian Naval Reserves", *Proceedings of the 2007 International Conference of the System Dynamics Society*, Boston, August.
- Moorhead, P. (2006) *A Strategic Intake Plan Model: Applied to the Pilot Occupation*, DRDC Technical Memorandum 2006-16, Centre for Operational Research and Analysis, Ottawa, Canada
- Straver, M., Okazawa, S., Wind, A., and Moorhead, P. (2009) *Training Pipeline Modelling Using the Production Management Tool*, DRDC Technical Memorandum 2009-019, Director General Military Personnel Research and Analysis, Ottawa, Canada.
- Straver, M. (2012) *Estimate of Required Throughput of Joint Command and Staff Program*, Briefing Note, Director General Military Personnel Research and Analysis, Ottawa, Canada.
- Straver, M. (2012) "Personnel Generation Performance Measurement in the Canadian Forces", *International Journal of Teaching and Education*, Vol 1, No. 6, pp 405-413.
- Straver, M., Latchman, S., and Tabbenor, N. (2013) "Estimating the Cost of a Proposed Change to Canadian Armed Forces Promotion Policy", *Proceedings of the 24<sup>th</sup> IASTED International Conference on Modelling and Simulation*, Banff, Canada, July.
- Wang, J. (2005) *A Review of Operations Research Applications in Workforce Planning and Potential Modeling of Military Training*, Defence Science and Technology Organisation Land Operations Division System Sciences Laboratory, DSTO-TR-1688, Salisbury, Australia.
- Woodill, G. (2005) *Forecasting Training Requirements for the Field Artillery*, DRDC Technical Memorandum 2005-11, Centre for Operational Research and Analysis, Ottawa, Canada.

# Detailed Maintenance Planning for Military Systems with Random Lead Times and Cannibalization

Rui Zhang and Ahmed Ghanmi

Center for Operational Research and Analysis, Defence Research and Development Canada, Ottawa Ontario, Canada

[Rui.Zhang@forces.gc.ca](mailto:Rui.Zhang@forces.gc.ca)

[Ahmed.Ghanmi@forces.gc.ca](mailto:Ahmed.Ghanmi@forces.gc.ca)

**Abstract:** Detailed maintenance planning under uncertainty is one of the most important topics in military research and practice. As one of the fastest ways to recover failed weapon systems, cannibalization operations are commonly applied by maintenance personnel. Due to the additional complexities introduced by these operations, detailed maintenance and repair decision making with cannibalization was rarely studied in the literature. This paper proposed an analytical model for studying the maintenance planning problem of military systems with random lead times and cannibalization. The objective of the problem is to maximize fleet reliabilities under operating costs constraints. A complementary problem that minimizes total operating costs under fleet reliabilities constraints was also constructed. A polynomial algorithm was proposed to solve the minimization problem and determine optimal decision strategies. This algorithm was also used as a subroutine in a binary-search algorithm to solve the maximization problem.

**Keywords:** Reliability, maintenance, optimization, cannibalization, multi-stage, multi-item

---

## 1. Introduction

### 1.1 Background

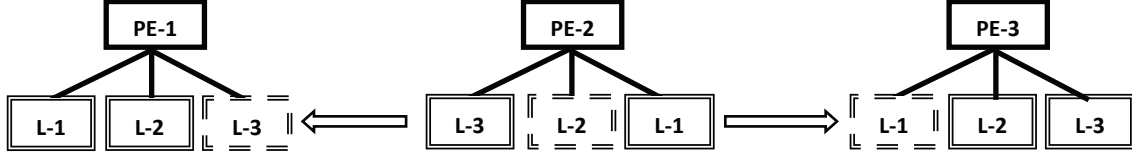
Manufactured products can fail due to different processes such as corrosion, wear and tear, as well as fatigue. Products such as domestic electronics and appliances are generally discarded and replaced upon failure because they are inexpensive. However, capital goods such as defence weapon systems are repaired because of their high replacement cost. Repair/maintenance decisions often involve removal and replacement of failed parts. Research on maintenance decision making problems have spawned several noteworthy papers in the open literature over the last five decades, starting with the textbook of Barlow and Proschan (1965) and the paper of McCall (1965). More recent research studies on a variety of models and solution methodologies for determining and/or comparing the best corrective, preventive and opportunistic maintenance policies could be found in the review papers (Dekker 1996), (Wang 2002) and (Nicolai and Dekker 2008), respectively.

In the military context, the research of maintenance systems focused on the level of repair analysis (LORA) problems (Basten et al. 2009), the spare parts stocking (SPS) problems (Tang and Liu 2011), and the combined LORA-SPS problems (Basten et al 2012). In these problems, repair decisions were studied at the strategic level using aggregated approaches. However, operating managers are under increased pressures to improve fleet reliabilities through their detailed (daily) repair decisions. This paper examined some of the repair decision making problems at the operational level and developed an optimization model to determine detailed maintenance planning strategies for military systems with random lead times and cannibalization.

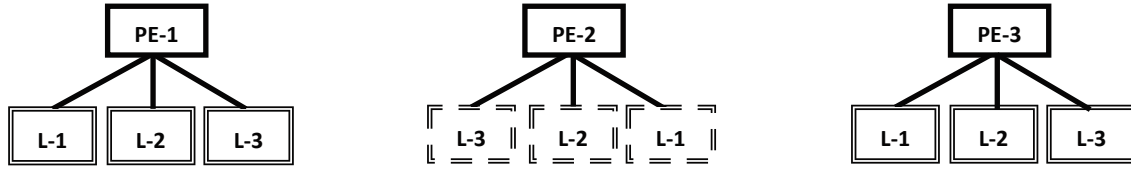
### 1.2 Cannibalization

In this study, a support network of one operating base and one repair depot was considered and a finite number of repair decision periods were used. In the network, when a defence weapon system or prime equipment (PE) fails, one and only one responsible part or line replaceable unit (LRU) is identified. The failed LRU will be separated from the PE and sent to the depot for repair. The left LRU hole will be filled up with a functioning LRU. If a spare LRU is not available, cannibalizing a functioning LRU from other failed PEs is also acceptable in practice. As indicated by the General Accounting Office of the United States (GAO U.S. 2001), the Air Force and the Navy reported 376,000 and 468,000 cannibalizations between 1996 and 2000, respectively. In the Canadian Armed Forces (CAF), the cannibalization of the Eryx missile system is being considered as the weapon will be retired in 2016.

Figures 1 and 2 describe an example of cannibalization, where the fleet includes three PEs, i.e., PE-1, PE-2, and PE-3, and each PE includes three LRUs, i.e., L-1, L-2 and L-3. In the figures, a functioning LRU is denoted by a solid-double-line rectangle; otherwise a dashed-double-line rectangle is used. Figure 1 has PE-1, PE-2 and PE-3 waiting for L-3, L-2 and L-1 replacement, respectively. The two arrows indicate that PE-1 and PE-3 can cannibalize L-3 and L-1 from PE-2, respectively. With these cannibalizations, PE-1 and PE-3 get back to work immediately, while PE-2 is in an even worse situation with three LRU “holes”. Figure 2 shows the resulting status of the fleet. It is clearly observed that the number of functioning PEs increases from zero (Figure 1) to two (Figure 2).



**Figure 1:** PE-1 and PE-3 can cannibalize L-3 and L-1 from PE-2, respectively.



**Figure 2:** PE-1 and PE-3 are recovered after cannibalization operations.

This example demonstrates that cannibalization is a useful operation when there is no sufficient spare LRU; this example also explains why cannibalization operations are favoured by military maintenance personnel in practice. However, due to the additional complexities introduced by cannibalization, detailed maintenance/repair decision making with cannibalization was rarely studied in the literature.

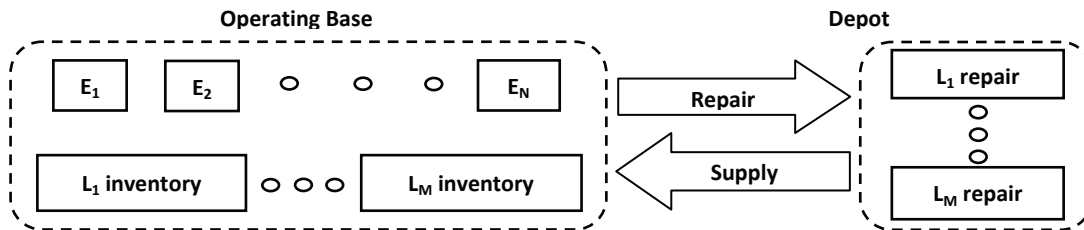
### 1.3 Structure

This paper is organized as follows. Section 2 describes the maintenance planning problem with random lead times and cannibalization. Section 3 presents a mathematical formulation of the maintenance planning problem and Section 4 proposes a binary-search algorithm to determine approximate solutions to the problem. Concluding remarks and future work are provided in Section 5.

## 2. Preliminaries

### 2.1 Maintenance system

A pictorial representation of the maintenance support network is shown in Figure 3. Initially, there are  $N$  ( $0 < N < \infty$ ) failed PEs ( $E_1, E_2, \dots, E_N$ ) installed in the operating base; each PE is made up of  $M$  ( $0 < M < \infty$ ) distinct LRUs ( $L_1, L_2, \dots, L_M$ ). Among the LRU positions, there is at least one malfunctioning LRU or LRU hole on each PE.



**Figure 3:** Maintenance support network with one operating base and one depot.

In the network, all failed PEs and spare LRUs are installed in the operating base and all repair operations are conducted in the depot (assuming that the warehouse space in the base and the repair capacity in the depot are unlimited).

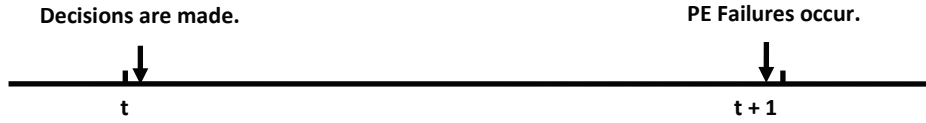
Let  $f_m$  and  $g_m$  be the number of functioning and malfunctioning  $L_m$  in the base, respectively. Let  $q_m$  be the number of such due-in  $L_m$ , which are either under repair in the depot or in transshipment between the base and depot. Thus, the total number of spare (or *individually existed*)  $L_m$  can be calculated as:

$$s_m = f_m + g_m + q_m - N.$$

The use of the term "individually existed" is due to the observation that by default a PE should include a PE frame, on which LRUs are installed. When the individually existed LRUs are first introduced in the system, they are not attached to any PE frame. Assuming that there are no external supplies or demands for any individual  $L_m$ ,  $s_m$  remains constant during the whole decision horizon.

A system is called *effective* if all failed PEs, whose  $L_m$  demands can be satisfied using the stocked  $L_m$  or the functioning  $L_m$  on other failed PEs (via cannibalization), are recovered. Given that all in-base operations are generally controlled by operating managers, the PE-recovery operations can be treated as zero-time and zero-cost operations. Assuming that the system is initially effective,  $N > 0$  implies that there is at least one  $L_m$  such that  $f_m = 0$ .

Let  $[1, T + 1)$ , where  $0 < T < \infty$ , be the decision horizon including  $T$  mutually disjoint time/decision periods. Figure 4 depicts an example of time period  $t$  ( $1 \leq t \leq T$ ), where repair decisions are made at the beginning of the period (i.e., just after time  $t$ ), and PE failures occur at the end of the period (i.e., just before time  $t + 1$ ).



**Figure 4:** Decisions are made at the beginning of period  $t$  and PE failures occur at the end of period  $t$ .

For a given decision period, let  $x_m$  be the decision variable denoting the number of  $L_m$  sent to the depot for repair and let  $k_m$  be the number of malfunctioning  $L_m$  introduced into the system by the PE failures, where  $x_m \geq 0$  and  $k_m \geq 0$  for all  $m = 1, 2, \dots, M$ . Based on the assumption that there is exactly one responsible LRU for each PE failure, the number of newly-failed PEs (or PE frames) is given by:

$$k = k_1 + k_2 + \dots + k_M.$$

The number of functioning  $L_m$  along with the newly-failed PEs is  $K_m = k - k_m$ . Therefore, the number of recovered PEs ( $n$ ) can be determined as follows (assuming zero spare  $L_m$  and zero repaired  $L_m$  returning/arriving during the period for all  $m = 1, 2, \dots, M$ ):

$$n = \min \{f_1 + K_1, f_2 + K_2, \dots, f_M + K_M\}.$$

Assuming that the system is effective, the recovered PEs will leave the system immediately. Thus, by the end of the period, only the number of unrecovered PEs is recorded. This paper assumes independent and identically distributed (IID) PE failures for all time periods.

## 2.2 Model definition

Let  $c^t$  be the total operating cost of period  $t$ , which is the sum of transportation (fixed and period-dependent) and repair (variable and LRU-dependent) costs. Let  $h^t$  be the transportation cost of period  $t$  if there is some malfunctioning LRUs sent to the depot for repair and let  $r_m$  be the per-unit repair cost for the made  $L_m$ -repair decisions. Thus,  $c^t = 0$  (if  $x_m = 0$  for all  $m$ ) and  $c^t = h^t + r_1x_1 + r_2x_2 + \dots + r_Mx_M$  (if there is at least one  $x_m > 0$ ). Let  $\gamma$  be the per-period interests gained on per unused fund. The present value of  $c^t$  at time is:  $pv^t = c^t / (1 + \gamma)^{t-1}$ , and the present value of the total operating cost over  $T$  periods is:

$$pv = pv^1 + pv^2 + \dots + pv^T.$$

On the other hand, let  $rc^t$  be the recovering ratio of the first  $t$  periods, which can be used to evaluate the fleet reliability of the first  $t$  periods ( $t = 1, 2, \dots, T$ ). Let  $k^t$  and  $n^t$  be the number of newly-failed and recovered PEs of period  $t$ , respectively. Then, the value of  $rc^t$  can be easily calculated as:

$$rc^t = \frac{\sum_{s=1}^t n^s}{N + \sum_{s=1}^t k^s}$$

Let  $B$  ( $0 < B < \infty$ ) be the total available operating budget at time one. The problem that maximizes fleet reliabilities with constraints on total operating costs (denoted by MaxRatio) can be formulated as:

$$\begin{aligned} \text{Max:} \quad & rc \\ \text{S.T.:} \quad & pv \leq B \\ & rc^t \geq rc, \text{ for all } t = 1, 2, \dots, T. \end{aligned}$$

In order to solve MaxRatio, a complementary problem is constructed to minimize total operating costs with constraints on fleet reliabilities. Similarly, this problem (denoted by MinCost) can be written as:

$$\begin{aligned} \text{Min:} \quad & pv \\ \text{S.T.:} \quad & rc^t \geq A, \text{ for all } t = 1, 2, \dots, T. \end{aligned}$$

Note that  $A$  ( $0 < A \leq 1$ ) is the designated level for fleet reliabilities.

### 3. Optimal solutions to MinCost

#### 3.1 Scenario tree formulation

The  $T$ -stage/period MinCost can be modelled on a  $(T+1)$ -level scenario tree, which is branched by PE failures. Let PE failures with small chances be zero-probability events. By assuming homogeneous Poisson processes (HPPs) for PE failures during each period, the maximum number of failed PEs during each period,  $K$  ( $0 < K < \infty$ ), can be easily determined for any given small chance. Therefore, the number of different PE failures during each period and the number of branches of the tree can also be determined. Note that in military maintenance problems, it is commonly adopted that IID-featured PE failures are described as HPPs.

Let  $W$  be the  $(T+1)$ -level scenario tree and  $V$  the set of nodes on  $W$ . Let  $l(j)$  be the level of node  $j$ , i.e.,  $1 \leq l(j) \leq T+1$  for all  $j$  in  $V$ . In particular,  $j$  is called a *leaf* node if  $l(j) = T+1$  and a *root* node if  $l(j) = 1$ . The index 1 is reserved for the root node;  $l(1) = 1$ . Let  $b(j)$  be the *direct* ancestor node of  $j$  and let  $B(j) = \{k | b(k) = j, \text{ for all } k \text{ in } V\}$  be the set of *direct* descendant nodes of  $j$ . It is obvious that  $b(1) = \emptyset$  and  $B(j) = \emptyset$  for all leaf node  $j$ . For a non-leaf node  $j$ , let  $W(j)$  be the sub-tree rooted on  $j$  and let  $V(j)$  be the set of nodes on  $W(j)$ . In particular,  $W(1) = W$  and  $V(1) = V$ .

Let  $p_j$  be the probability of node  $j$ . Since node 1 represents the system at time one, where the status is well-defined, then  $p_1 = 1$ . Since all nodes in  $B(1)$  are branched from node 1, the sum of probabilities of nodes in  $B(1)$  is equal to one (the sum of probabilities of all leaf nodes is also equal to one).

For a node  $j$ , let  $P(j)$  be the path including all nodes from 1 to  $j$ . In particular,  $P(1) = \{1\}$ . If  $j$  is a leaf node, then  $P(j)$  is called a scenario. Note that a scenario is a series of events, and the probability of a scenario is represented by the probability of the leaf node, with which the scenario ends up. All reliability and cost calculations are based on either paths or scenarios.

In order to formulate MinCost on  $W$ , fleet reliability requirements are converted to LRU demands first. On  $W$ , let  $(k, j)$  be the edge connecting node  $k$  and  $j$ . Obviously,  $(k, j)$  denotes the PE failures occurring in period  $l(k)$ ,



$l(j)$ ). Let  $D_{(j,m)}$  be the cumulative demands for  $L_m$  by node  $j$ . Let  $k_j$  be the number of failed PEs introduced by  $(k,j)$  and let  $k_{(j,m)}$  be the number of malfunctioning  $L_m$  associated with  $(k,j)$ . For any give node  $j > 1$ ,  $D_{(j,m)}$  can be easily determined as (assuming  $A = 1$  and  $s_m = 0$  for all  $m$ ):

$$D_{(j,m)} = \max_{m=1,2,\dots,M} \{0, N + \sum_{i \in P(j)} k - f_{(1,m)} - \sum_{i \in P(j)} K_{(i,m)}\}.$$

Since PE failures always bring functioning LRUs and these LRUs can be used as spare LRUs, there might be some irregular cumulative demands, i.e.,  $D_{(j,m)} < D_{(b(j),m)}$  for some  $j$  and  $m$ . In this case,  $D_{(j,m)}$  is re-assigned to  $D_{(b(j),m)}$ .

Let  $x_{(j,m)}$  be the decision variable, which represents the number of  $L_m$  sending to the depot for repair on node  $j$ . Let  $y_j$  be the binary variable such that  $y_j = 1$  if there is some  $m$  with  $x_{(j,m)} > 0$  and  $y_j = 0$  if  $x_{(j,m)} = 0$  for all  $m$ . Let  $d_{(j,m)} > 0$  be the lead time for the  $L_m$ -repair operations on node  $j$ . Due the fact that none out-base operations are controlled by operating managers, this paper assumes random  $d_{(j,m)}$  variables. Note that  $d_{(j,m)}$  are assumed to be positive integers. Let  $F(j,m) = \{k | l(k) = l(j) + d_{(j,m)} \text{ and } k \text{ in } V(j)\}$  be the set of nodes such that if a failed  $L_m$  is sent out for repair on node  $j$ , then the repaired  $L_m$  will be returned on all nodes in  $F(j,m)$ . On the other hand, let  $G(j,m) = \{k | l(k) = l(j) + d_{(j,m)} \text{ and } k \text{ in } P(j)\}$  be the set of nodes such that if a failed  $L_m$  is sent out for repair on any node in  $G(j,m)$ , then the repaired  $L_m$  will be returned on node  $j$ . Next, the scenario tree formulation of MinCost, denoted by MinCostST, is formulated as follows:

$$\text{Min: } \sum_{j \in V: l(j) < T+1} p_j \frac{h_j y_j + \sum_{m=1}^M r_m x_{(j,m)}}{(1 + \gamma)^{l(j)-1}} \quad (1)$$

$$\text{S.T.: } D_{(j,m)} \leq \sum_{i \in P(j): j \in V(h) \text{ where } h \in F(i,m)} x_{(i,m)} \text{ for all } j, m \quad (2)$$

$$x_{(j,m)} \leq g_{(1,m)} + \sum_{i \in P(j)/\{1\}} (k_{(i,m)} - x_{(b(i),m)}) \text{ for all } j, m \quad (3)$$

$$x_{(j,m)} \leq y_j Z \text{ for all } j, m \quad (4)$$

$$x_{(j,m)} \text{ are non-negative integers for all } j, m \quad (5)$$

$$y_j \text{ are binary variables for all } j. \quad (6)$$

Equation (1) is the objective to minimize the expected present value of the total costs. Constraints (2) and (3) represent the demand satisfactions and LRU-repair capacities, respectively. Constraint (4) uses the large  $Z$  value to model the relationship between  $x_{(j,m)}$  and  $y_j$ . This makes MinCostST a non-trivial stochastic, production planning problem (PPP) with dynamically varied production capacities.

### 3.2 Polynomial algorithm

A deterministic PPP has all demands well-defined before time one. Using the Wagner-Whitin property (Wagner and Whitin 1958), if the problem has zero initial inventory, zero lead times and unlimited ordering capacities, then there exists an optimal plan such that the accumulated order quantity by period  $t$  is exactly the accumulated demand by some period  $s$ , where  $0 < t < s < \infty$ . However, with additional stochastic features, i.e., demands will be available only when times come, the Wagner-Whitin property is invalid (Ahmed 2003). In order to deal with this stochastic PPP, the Production-Path property was proved in (Guan and Miller 2008). In (Huang and Kucukyavuz 2008), the Production-Path property was modified to solve a more general stochastic PPP, where all lead times, production and inventory costs were random. Note that these PPPs consider only one type of product/item and the production/ordering capacities are unlimited. In MinCostST, however, each PE is made up of  $M$  LRUs and  $L_m$ -repair operations are upper-bounded by the number of malfunctioning  $L_m$ . The following proposition modifies the above properties to cooperate these stochastic features in MinCostST.

**Proposition 1** For MinCostST, there exists an optimal solution  $\Omega = \{x_{(j,m)}, y_j | \text{ for all } j \text{ in } V\}$  such that if  $x_{(j,m)} > 0$  for some  $j$  and  $m$ , then (I)  $y_j = 1$ ; (II)  $x_{(j,m)}$  satisfies the repair-capacity constraint (3); and (III) the demand-satisfaction constraint (2) is addressed by:

$$\sum_{k \in G(i,m) \cap P(j)} x_{(k,m)} = D_{(h,m)} - D_{(b(i),m)}, \text{ where } h \in V(i) \text{ and } i \in F(j,m).$$

*Proof* This proposition is proved by contradiction. An equivalent statement for this proposition is that for node  $j$  either there is no any LRU-repair operation or each LRU-repair operations are used to exactly satisfy the demands on some nodes at which the repaired LRUs will arrive later.

For  $L_m$ , let node  $j$  be the node with the smallest level such that a positive  $x_{(j,m)}$  is not used to exactly satisfy the  $L_m$  demand on any node in  $V(k)$ , where  $k$  could be any node in  $F(j,m)$ . Note that  $V(k)$  is in  $V(j)$ .  $\Omega$  is optimal implies that all  $L_m$  demands on all nodes in  $V$  are satisfied by the current repair decisions in  $\Omega$ . Reducing the value of  $x_{(j,m)}$  to its minimum value such that any further reduction on  $x_{(j,m)}$  would cause unsatisfied  $L_m$  demand on some node in  $V(k)$  for some  $k$  in  $F(j,m)$ . In this case, there is at least one node in  $V(j)$ , whose  $L_m$  demand is satisfied either by  $x_{(j,m)}$  only or by a combination of  $x_{(j,m)}$  and some later  $L_m$ -repair operations. In either case, contradiction is found as the conducted reduction on  $x_{(j,m)}$  would not cause any increases on operating costs, but make the proposition satisfied.

Proposition 1 indicates that a dynamic programming algorithm, which evaluates possible LRU repair decisions on all non-leaf nodes in  $V$ , can be developed to search for optimal solutions to MinCostST. Let  $v_j(\alpha)$ , where  $\alpha = (\alpha(1), \alpha(2), \dots, \alpha(M))$ , be the optimal solution of the MinCostST problem defined on  $W(j)$ , where  $D_{(\alpha(m),m)}$  is satisfied by the  $L_m$ -repair decisions made on  $P(b(j))$  for all nodes  $\alpha(m)$  in  $\alpha$ . Note that  $\alpha$  might denote a single node (if  $\alpha(1) = \alpha(2) = \dots = \alpha(M)$ ) or at most  $M$  distinct nodes (if  $\alpha(j) \neq \alpha(k)$  for any  $j$  and  $k$  with  $1 \leq j, k \leq M$  and  $j \neq k$ ). Let  $u_j(\alpha, \beta)$ , where  $\alpha = (\alpha(1), \alpha(2), \dots, \alpha(M))$  and  $\beta = (\beta(1), \beta(2), \dots, \beta(M))$ , be the objective value of the MinCostST problem defined on  $W(j)$ , where  $D_{(\alpha(m),m)}$  are satisfied by the  $L_m$ -repair decisions made on  $P(b(j))$  for all  $\alpha(m)$  in  $\alpha$  and the LRU-repair decisions on node  $j$  would be either  $x_{(j,m)} = D_{(\beta(m),m)} - D_{(\alpha(m),m)}$  or  $x_{(j,m)} = 0$ . Considering a positive  $x_{(j,m)}$ , the upper bound of  $x_{(j,m)}$  can be calculated as:

$$g(j, m) = g_{(1,m)} + \sum_{i \in P(j)} (k_{(i,m)} - x_{(b(i),m)}).$$

Thus, the value of  $u_j(\alpha, \beta)$  can be calculated as:

$$u_j(\alpha, \beta) = \begin{cases} p_j \frac{h_j + \sum_{m=1}^M x_{(j,m)} r_m}{(1+\gamma)^{l(j)-1}} + \sum_{k \in B(j)} v_k(\alpha), & \text{if for some } m \\ & \text{it has } 0 < D_{(\beta(m),m)} - D_{(\alpha(m),m)} \leq g_{(j,m)} \\ \sum_{k \in B(j)} v_k(\alpha), & \text{if it has } D_{(\beta(m),m)} - D_{(\alpha(m),m)} = 0 \\ & \text{or } D_{(\beta(m),m)} - D_{(\alpha(m),m)} > g_{(j,m)} \text{ for all } m \end{cases}. \quad (7)$$

Note that  $v_k(\alpha)$  is the optimal solution for the MinCostST problem defined on  $W(k)$ , where  $D_{(\alpha(m),m)}$  are satisfied by the  $L_m$ -repair decisions on  $P(b(k))$  for all  $\alpha(m)$  in  $\alpha$ . It can be calculated as:

$$v_k(\alpha) = \min_{\beta(m) \in \{h|h \in V(k) \text{ and } l(h) \geq l(k) + d_{(k,m)}\}: D_{(\beta(m),m)} - D_{(\alpha(m),m)} \geq 0} \{u_k(\alpha, \beta)\}. \quad (8)$$

The following backwards dynamic programming algorithm, denoted by Recursion, starts from  $v_1(\bullet)$ , where  $\bullet$  is the dummy notation for the initial system status. Regarding boundary conditions, for node  $j$  in  $V$ , if  $l(j) + d_{(j,m)} > T+1$ , then the optimal  $L_m$ -repair decision is  $x_{(j,m)} = 0$ . Thus, if  $l(\beta(m)) + d_{(\beta(m),m)} > T+1$  occurs for all  $\beta(m)$  in  $\beta$ , then  $u_j(\alpha, \beta) = 0$  is determined for all feasible  $\alpha$ .

#### Recursion

[Boundary Conditions]: For any pair of  $(\alpha, \beta)$ , if  $l(\beta(m)) + d_{(\beta(m),m)} > T+1$  for  $m$  and all  $\beta(m)$  in  $\beta$ , then set  $u_j(\alpha, \beta) = 0$ ; otherwise set the value of  $u_j(\alpha, \beta)$  as undetermined.

[Recursion Procedure]: For each  $u_j(\alpha, \beta)$ , perform the calculations as described in Equation (7).  
 For each  $k$  in  $B(j)$ , perform the calculations as described in Equation (8).  
 [Optimal Solution]: Perform the calculations described in Equation (8), where  $v_k(\alpha)$  and  $u_k(\alpha, \beta)$  are replaced by  $v_1(\bullet)$  and  $u_1(\bullet, \beta)$ , respectively. (Note that the assumption of all non-zero lead times gives  $D_{(1,m)} = 0$  for all  $m$ .)

**Theorem 1** Recursion can find an optimal solution to MinCostST in  $O(|V|^{2M+2})$  time, where  $|V|$  denotes the total number of nodes in  $V$ .

*Proof* The correctness of this theorem follows Proposition 1 and the discussions above. The total number of  $u_j(\alpha, \beta)$  is upper-bounded by the production of the number of  $j$ ,  $\alpha$  and  $\beta$ , whose maximum values are  $|V|$ ,  $|V|^M$  and  $|V|^M$ , respectively. Since each  $u_j(\alpha, \beta)$  value requires to evaluate exactly  $|B(j)|$  number of  $v_k(\alpha)$ , i.e., one evaluation for each node  $k$  in  $B(j)$ , the total calculations require at most  $O(|V|^{2M+2})$  such evaluations, where  $|B(j)| < |V|$  is the total number of nodes in  $B(j)$ . Since computing  $u_j(\alpha, \beta)$  dominates the run time of Recursion, the overall run time is  $O(|V|^{2M+2})$ .

#### 4. Approximation solutions to MaxRatio

In MaxRatio, the goal is to determine a set of repair operations to maximize fleet reliabilities with constraints on operating costs. In military maintenance problems, however, if repair operations have direct/significant effects on fleet reliabilities, then there won't be any strict limit on operating costs. For example, in the CAF, *contingency funds* are always available for unexpected repair operations as long as there is a need. Thus, in the scenario tree formulation of MaxRatio (denoted by MaxRatioST), the *expected* value of the total  $T$ -period operating costs is constrained by the limited budget.

Let  $D_{(j,m)}(r)$  denote the  $L_m$  demand by node  $j$  in  $V$  with respect to required fleet reliability level  $r$ , where  $m = 1, 2, \dots, M$  and  $0 \leq r \leq 1$ . Note that the value of  $D_{(j,m)}$  used in MinCostST is based on the fleet reliability level  $A$ , i.e.,  $D_{(j,m)} = D_{(j,m)}(A)$ . As discussed in Section 3.1, for a given  $r$ ,  $D_{(j,m)}(r)$  can be easily determined. However, if  $s_m > 0$  for some  $m$  or  $r < 1$ , then there would be no explicit formula to calculate  $D_{(j,m)}(r)$ . This makes MaxRatioST, which includes  $D_{(j,m)}(r)$  as parts of demand satisfaction constraints, mathematically intractable. The MaxRatioST problem is formulated as follows:

$$\text{Max: } r \quad (9)$$

$$\text{S.T.: } D_{(j,m)}(r) \leq \sum_{i \in P(a(j)); j \in V(k) \text{ where } k \in G(i,m)} x_{(i,m)} \text{ for all } j, m \quad (10)$$

$$\sum_{j \in V} p_j \frac{h_j y_j + \sum_{m=1}^M r_m x_{(j,m)}}{(1 + \gamma)^{l(j)-1}} \leq B \quad (11)$$

$$x_{(j,m)} \leq g_{(1,m)} + \sum_{i \in P(j)} (k_{(i,m)} - x_{(b(i),m)}) \text{ for all } j, m \quad (12)$$

$$x_{(j,m)} \leq y_j Z \text{ for all } j, m \quad (13)$$

$$0 \leq r \leq 1 \quad (14)$$

$$x_{(j,m)} \text{ are non-negative integers for all } j, m \quad (15)$$

$$y_j \text{ are binary variables for all } j. \quad (16)$$

Equation (9) indicates that the objective of the problem is to maximize the fleet reliability level and Equation (14) bounds  $r$  into a feasible range  $[0,1]$ . Constraints (10) and (11) describe the demand and the cost constraints, respectively. All other constraints (12), (13), (14) and (15) have the same meanings as in MinCostST.

Since MaxRhoSTM includes some implicit terms, it is unlikely that MaxRhoSTM can be directly solved using general optimization approaches, e.g., mathematical programming. In this case, combinatorial optimization approaches (e.g., dynamic programming) might be useful tools for optimal solutions. Generally, these approaches enumerate and compare all possible solutions to determine the optimal solution. This requires a finite number of value options for decision variables. In MaxRhoSTM, however,  $r$  could take any real value in

[0,1] and the number of value options for  $r$  is infinity. It is impossible to develop an optimal algorithm for MaxRatioSTM using general combinatorial optimization approaches.

Next, a solution algorithm (BinRec) is developed to determine approximation solutions to MaxRatioST. In BinRec, a binary search is performed. In each iteration, an on-hand  $r$  value is first used to determine  $D_{(j,m)}(r)$  and  $\text{MinCostST}(r)$ , and then Recursion is called to solve  $\text{MinCostST}(r)$ . If Recursion returns an *acceptable* result (i.e., a feasible solution is found to  $\text{MinCostST}(r)$ , where the objective value is within the budget limit), then  $r$  is updated to a larger value for the next iteration. If Recursion returns no *acceptable* result (i.e., either none feasible solution is found to  $\text{MinCostST}(r)$  or the found feasible solution has the objective value over the budget limit), then  $r$  is updated to a smaller value for the next iteration.

Before presenting BinRec, let  $r^*$  and  $r(\epsilon)$  be the optimal and approximation solution values of MaxRatioST, respectively. It is obvious that  $r(\epsilon) \leq r^*$ , e.g.,  $r(\epsilon) \geq r^* - \epsilon$  for some  $\epsilon > 0$ . In BinRec, let  $U$  and  $L$  be the upper and lower bounds of  $r^*$ , respectively. Let  $e = U - L$  be the bound range. Assuming that there is at least one feasible solution to MaxRatioST, at least one failed PE can be recovered under the budget limit. Thus, the initial lower bound is  $L = 1/(TK+N)$ . Note that  $K$  denotes the maximum for the number of failed PEs introduced in each period. It is also true that the initial upper bound is  $U = 1$ . Therefore, the initial range is determined as  $e = U - L = 1 - 1/(TK+N)$ .

#### BinRec

[Preparation]: Set  $L = 1/(TK+N)$ ,  $U = 1$ ,  $e = 1 - 1/(TK+N)$ .  
 Set  $r(\epsilon) = L$ , where  $\epsilon$  is selected as  $0 < \epsilon < e$ .  
 [BinarySearch]: If  $e \leq \epsilon$ , then go to [FinalCheck]. Otherwise,  
 • Set  $r = (L + U)/2$  and determine  $D_{(j,m)}(r)$ .  
 • Run Recursion to solve  $\text{MinCostST}(r)$ .  
     ○ If there are no acceptable solution, then set  $U = r$ .  
     ○ If there is an acceptable solution, then set  $L = r$  and  $r(\epsilon) = L$ .  
 • Calculate  $e = U - L$  and go to [BinarySearch].  
 [FinalCheck]: Determine  $D_{(j,m)}(U)$ , and run Recursion to solve  $\text{MinCostST}(U)$ .  
 • If there is no acceptable result, then go to [solution].  
 • If there is an acceptable result, then set  $r(\epsilon) = U$ .  
 [Solution]: Select  $r(\epsilon)$  as the found fleet reliability level.  
 Trace back to obtain the corresponding repair decisions.

**Theorem 2** For MaxRatioST, BinRec can find an approximation solution  $r(\epsilon)$  such that  $r(\epsilon) > r^* - \epsilon$  in  $O([\ln(TK+N-1)/(\epsilon TK + \epsilon N)]|V|^{2M+2})$  time.

*Proof* In [BinarySearch], the bounds are cut into half after each iteration by either increasing  $L$  or decreasing  $U$ .  $r(\epsilon)$  is updated when a feasible solution is found. After [FinalCheck], it is determined that  $r^*$  in  $[L, U]$  with  $L = r(\epsilon)$ . The stopping criteria ( $e \leq \epsilon$ ) gives  $r^* < U \leq L + \epsilon = r(\epsilon) + \epsilon$ , i.e.,  $r(\epsilon) > r^* - \epsilon$ . For complexity, in each iteration Recursion runs in  $O(|V|^{2M+2})$  time to solve  $\text{MinCostST}(r)$ . Since BinRec runs a binary search, Recursion is called  $O([\ln(TK+N-1)/(\epsilon TK + \epsilon N)])$  times. Thus, the overall run time is  $O([\ln(TK+N-1)/(\epsilon TK + \epsilon N)]|V|^{2M+2})$ . Note that solution accuracy is traded-off with run time.

Note that BinRec has controllable run times and approximation ratios such that a pre-determined value of  $\epsilon$  can be used to estimate the solution quality (i.e.,  $r(\epsilon) > r^* - \epsilon$ ) and the run time (i.e.,  $O([\ln(TK+N-1)/(\epsilon TK + \epsilon N)]|V|^{2M+2})$ ). This demonstrates that more accurate solutions require longer run times. As an extreme case, however,  $\epsilon = 0$  requires unlimited run time. This shows that BinRec cannot be used to search for optimal solutions to MaxRatioST.

## 5. Final remarks

This paper proposed an analytical model for multi-stage repair/maintenance decision making in a single-base, single-depot military support network, where cannibalization operations were allowed; failures were independent; repair lead times were random. The paper presented a solution algorithm for the problem to minimize total operating costs and an approximation algorithm for the problem to maximize fleet reliabilities. Future research could be to extend this study to more complex maintenance systems, where support networks

are structured by several operating bases and/or repair depots. Another research direction could be to develop optimal algorithms for the fleet-reliability-maximization problem.

## **References**

- Zhang, R. (2013) Detailed Cannibalization Decision Making for Maintenance Systems in the Military Context, Technique Memorandum 2013-228, DRDC-CORA, Ottawa, Canada.
- Barlow, R. and Proschan, F. (1965) Mathematical Theory of Reliability, John Wiley & Sons, New York.
- McCall, J. (1965) "Maintenance Policies for Stochastically Failing Equipment: A Survey", *Management Science*, Vol 11, No. 5, pp 493–524.
- Dekker, R. (1996) "Applications of Maintenance Optimization Models: A Review and Analysis", *Reliability Engineering & System Safety*, Vol 51, No. 3, pp 229–240.
- Wang, H. (2002) "A Survey of Maintenance Policies of Deteriorating Systems" *European Journal of Operational Research*, Vol 139, No. 3, pp 469–489.
- Nicolai, R. and Dekker, R. (2008) Optimal Maintenance of Multi-component Systems: A Review, Complex System Maintenance Handbook, Springer Series in Reliability Engineering, pp. 263–286.
- Basten, R., Schutten, J., and der Heijden, M. (2009) "An Efficient Model Formulation for Level of Repair Analysis", *Annals of Operations Research*, Vol 172, No. 1, pp 119–142.
- Tang, L. and Liu, X. (2011) "Spares Provisioning for Repairable Systems under Fleet Expansion", *IEEE Proceedings on 2011 Annual Symposium - RAMS*, pp 1–5.
- Basten, R., der Heijden, M. and Schutten, J. (2012) "Joint Optimization of Level of Repair Analysis and Spare Parts Stocks", *European Journal of Operational Research*, Vol 222, No. 3, pp 474–483.
- GAO U.S. (2001), Military Aircraft Services Need Strategies to Reduce Cannibalizations. GAO-02-86.
- Wagner, H. and Whitin, T. (1958) "Dynamic Version of the Economic Lot Size Model", *Management Science*, Vol 5, No. 1, pp 89–96.
- Ahmed, S., King, A., and Parija, G. (2003) "A Multi-stage Stochastic Integer Programming Approach for Capacity Expansion under Uncertainty", *Journal of Global Optimization*, Vol 26, pp 3–24.
- Guan, Y. and Miller, A. (2008) "Polynomial-Time Algorithms for Stochastic Uncapacitated Lot-Sizing Problems", *Operations Research*, Vol 56, No. 5, pp 1172–1183.
- Huang, K. and Kucukyavuz, S. (2008) "On Stochastic Lot-Sizing Problems with Random Lead Times", *Operations Research Letters*, Vol 36, pp 303–308.

# **Work in Progress Papers**





# Knowledge-Based Cause-Effect Analysis Enriched by Generating Multi-Layered DSS Models

Axel Benjamins

University of Osnabrueck, Germany

[abenjamins@uos.de](mailto:abenjamins@uos.de)

**Abstract:** Computer-based analysis to support decision-making in organizations is a crucial competitive factor. Cause-effect analysis is an important component of these analyses, as it identifies cause-effect relationships amongst data, which can be applied in decision-making situations to improve the decision-making quality. This paper envisions a concept for the support of cause-effect analyses, which is based on an integrated knowledge base with cause-effect relationships and a knowledge reasoning process according to the human approach to solving problems. The knowledge base integrates both structured and unstructured knowledge from a variety of organizational sources. The knowledge reasoning is divided into three phases during which the decision situation is (1) isolated and matched into the knowledge base, (2) explored for potential causes (including their validation) and finally (3) verified, and, if necessary, adjusted by the user. As a proof of concept, this concept is applied manually to the slightly extended example data set from Microsoft for the SQL Server 2012. For the creation of the knowledge base, knowledge about the cause-effect relationships is extracted manually from the database schemas and integrated with additional expert knowledge about further cause-effect relationships. The result is an ontology with cause-effect relationships for this specific data set. Based on a fictitious decision scenario the phases of the knowledge reasoning are played through. The exploration of the ontology will typically identify cause-effect chains with various potential explanations alongside the levels of the chain. These potential cause-effect chains are implemented in a DSS model with multiple layers. The resulting DSS model enables the evaluation of the impact of the identified cause-effect chains for the specific decision scenario.

**Keywords:** cause-effect analysis, knowledge-based, DSS model, decision support

---

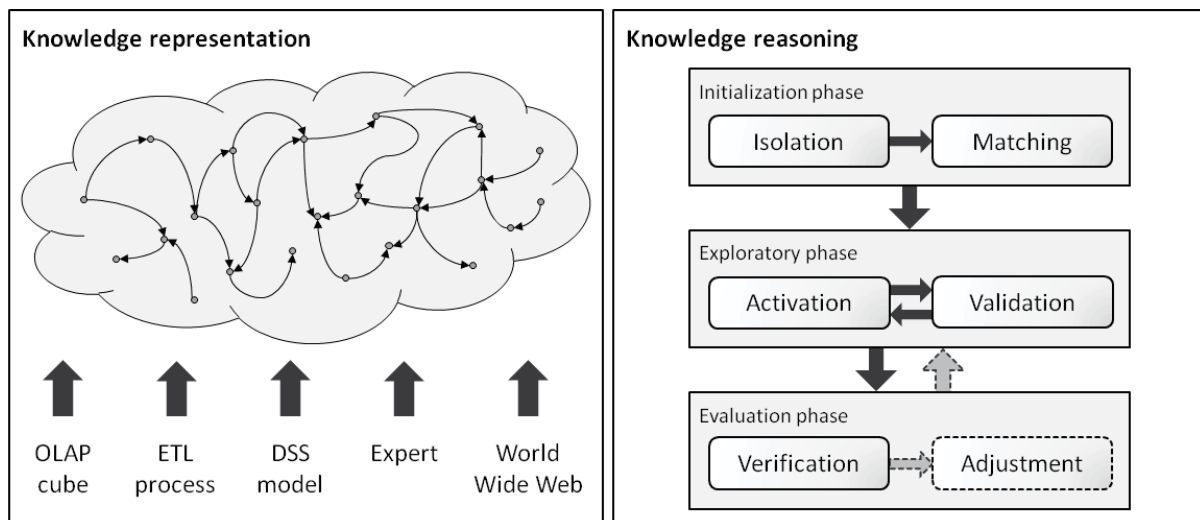
## 1. Introduction

Companies that apply analytics to gain a competitive advantage (Davenport & Harris 2007, p.48) can make more effective and timely decisions (Vercellis 2009, p.5). Cause-effect analysis is an important component in these analyses (Mintzberg et al. 1976, p.274), as it clarifies and defines a decision-making situation (Mintzberg et al. 1976, p.254) through the identification of potential causes for a specific effect. The impending shortage of employees with analytical competencies (Chen et al. 2012, p.1185) and the importance of cause-effect analyses for the understanding of a decision-making situation leads to the conclusion that it might be necessary to offer computer-supported cause-effect analysis to all decision makers to sustain a competitive advantage.

This paper demonstrates the feasibility of the concept for a knowledge-based cause-effect analysis from Benjamins (2014). The concept is manually applied to an example data set to support a fictitious decision scenario by generating a multi-layered decision support system (DSS) model.

## 2. Concept

The following is an overview of the concept for a knowledge-based cause-effect analysis. A detailed description can be found in Benjamins (2014). The concept uses the separation of the knowledge representation and the knowledge reasoning, which is typical for a knowledge-based system (Brachman et al. 2004).



**Figure 1:** Overview of a knowledge-based cause-effect analysis (adapted from Benjamins (2014))

For the knowledge representation, knowledge about cause-effect relationships is extracted from a variety of sources within a company, transformed into a unified structure and loaded into the knowledge base. The unified structure consists of knowledge elements, relationships between these elements and different relationship types. The sources can contain structured and unstructured knowledge. Foremost, formulas of DSS models provide structured knowledge about quantified relationships between data. OLAP cubes also implicate potential cause-effect relationships within their dimensional structure and even ETL processes from a Data Warehouse implicitly contain cause-effect knowledge within their transformations. Linked (Open) Data from the World Wide Web can add external influences. This is enriched with unstructured knowledge about cause-effect relationships from experts to add relationships between data from different sources. The result is a homogeneous knowledge base with knowledge about cause-effect relationships from a variety of sources (ref. Figure 1).

The knowledge reasoning is applied to the knowledge base in order to identify relevant causes in a specific decision-making situation. The reasoning is based on the separation of a decision-making process into the three phases intelligence, design and choice (Simon 1977, p.40) and the human approach to solve problems by using a specific problem context (Newell & Simon 1972, p.809). The reasoning is divided into three phases with two steps each (ref. Figure 1). During the initialization phase, a decision-making situation is isolated and matched into the knowledge base. This identifies the relevant factors for the decision-making situation within the knowledge base. The exploratory phase uses these identified factors in the activation and validation steps. During activation, all relations between elements directly connected to the identified factors are marked as unconfirmed cause-effect relationships. These unconfirmed relationships are statistically validated with the help of time-series data, e.g. from data warehouse or operational systems. If they are confirmed, the connected elements are promoted to relevant factors and used for another activation step. The result is a targeted activation along a chain of promising factors in a specific decision-making situation. The evaluation phase offers the user the possibility to verify the confirmed cause-effect relationships and, if necessary, to adjust the affected relationships in the knowledge base.

### 3. Proof of concept

The feasibility of the concept is proven by manually applying it to an example data set. In the following chapter, the selected data set is described, the knowledge representation is presented and the knowledge reasoning played through for a specific decision scenario. Additionally, the generated DSS model is explained.

#### 3.1 Example data set

The sample database for the fictitious company Adventure Works Cycles is available for all major versions of the Microsoft SQL Server (Microsoft 2014). Adventure Works Cycles is a global company that manufactures and sells bicycles. The manufacturing is done in Mexico and the base operation is in the USA with additional regional sales teams. There are sample databases for an online transaction processing (OLTP) database and a

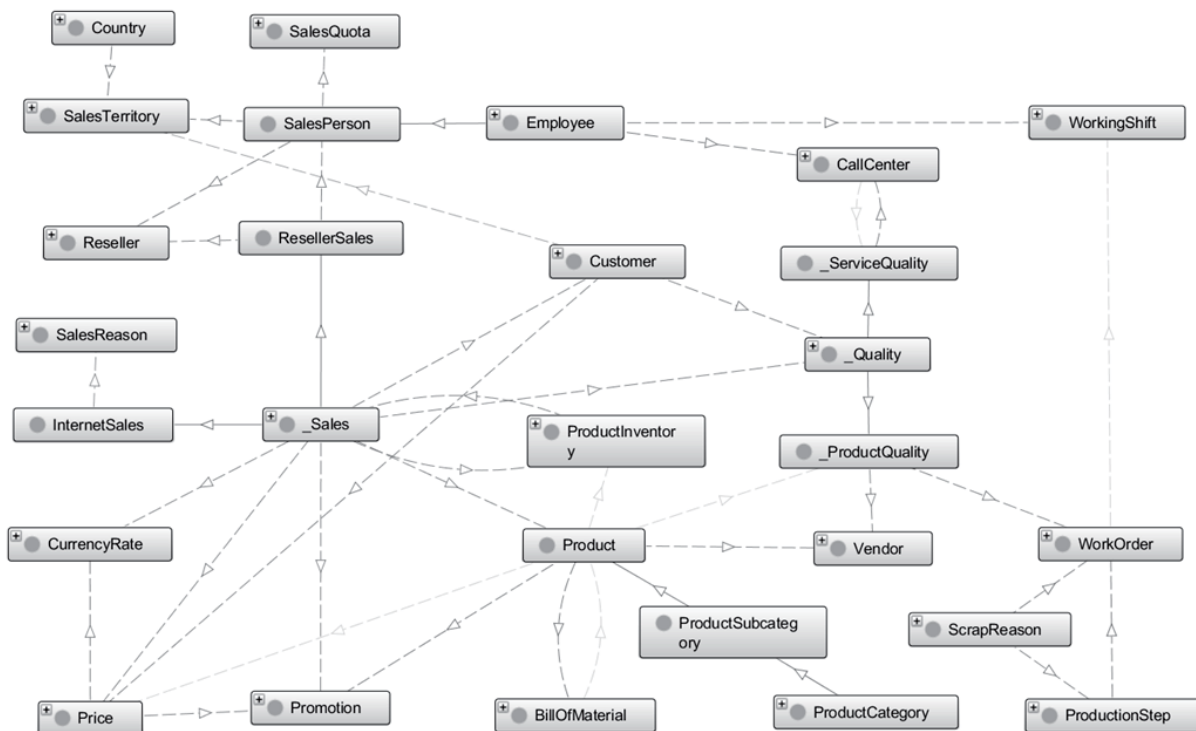
data warehouse (DW). The OLTP database contains data about *Human Resources, Purchasing, Production, Persons* and *Sales*. The DW includes the subjects *finance* and *sales*, with multiple fact tables each. For this proof of concept, the sample databases for the OLTP as well as the DW are used in the versions for the Microsoft SQL Server 2012. The OLTP database was slightly extended with a table about working shifts for the manufacturing process.

Based on this example data set, a fictitious scenario of dropped sales in a specific product subcategory in a specific sales territory will be analyzed to discover possible causes.

### 3.2 Knowledge representation

Knowledge about cause-effect relationships is extracted from the OLTP and DW data model, transferred into a unified structure, and thus, a knowledge base is created. This is done by representing a database table as an ontology class and important table columns as annotation to a class (Gómez-Pérez et al. 2003, p.23). Additionally, a relationship between two tables in the data model is represented as a connection between the two correspondent classes (Gómez-Pérez et al. 2003, p.23), e.g. a connection between the fact table *FactResellerSales* and the dimension table *DimReseller* depicts that reseller sales are potentially influenced by the reseller or vice versa. The direction of the influence must be defined by an expert. The data and relationships are then checked for duplicates and possibilities to unify nodes. The result is used for the creation of an OWL ontology (W3C OWL Working Group 2009) to represent the knowledge about (types of) cause-effect relationships (and corresponding impact calculations) based on these data models (ref).

Figure 2, different relationship types, e.g. solid versus dashed lines, represents different impact calculations. The impact calculations are used to compute the correlation between two related factors.



**Figure 2:** Ontology representing knowledge about cause-effect relationships for the AdventureWorks database

The ontology is extended with abstract nodes, which don't originally belong to the data models, but are added by a human expert, e.g. *\_Quality* is non-existing in the data models, but does potentially have an influence on sales (especially *\_ProductQuality*).

### 3.3 Knowledge reasoning

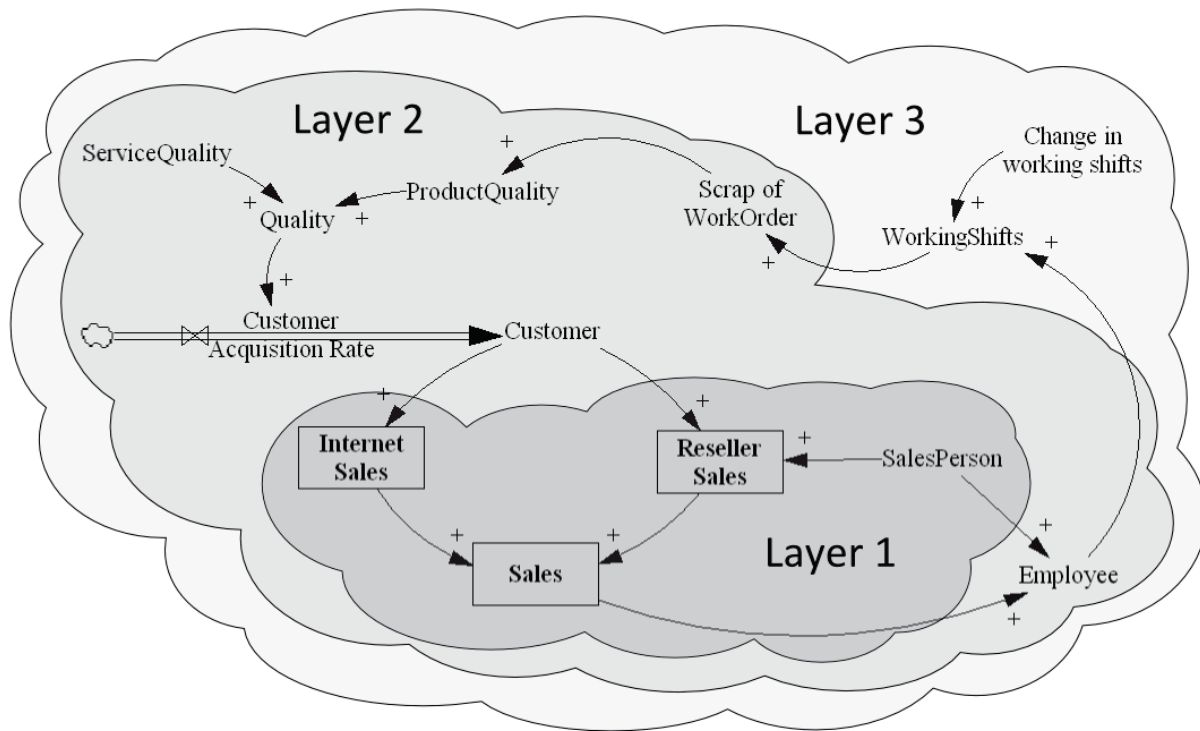
The knowledge reasoning (ref. Chapter 2) is manually applied to the created knowledge base (ref. Chapter 3.2) for the fictitious scenario (ref. Chapter 3.1), thus showing the procedure of the reasoning phases initialization, exploration and evaluation.

The initialization phase maps the scenario into the knowledge base. First, the factors of the fictitious scenario of *dropped sales* are identified (isolated), which are *sales*, *product subcategory* and *sales territory*. Then, the knowledge base is searched for these identified factors, which are matched with appropriate elements. The matched elements *\_Sales*, *SalesTerritory* and *ProductSubcategory* are used as starting points for the exploration iterations. All elements connected to these starting points are activated for further analysis, e.g. the factor *\_Sales* activates the elements *InternetSales*, *ResellerSales*, *Customer*, *\_Quality*, *ProductInventory*, *Product*, *Promotion*, *Price* and *CurrencyRate* (ref. Figure 2). The relevance of the activated elements in the specific scenario is validated by calculating the impact of an activated element on the initial factor, e.g. the impact of *Customer* on *\_Sales*. The calculation for the impact depends on the relationship type between the two elements, e.g. the impact of *Customer* on *\_Sales* is calculated with the Pearson correlation, because of the relationship of the type *is\_influenced\_by*. The data for the calculation is fetched from the OLTP or DW databases. An activated element is successfully validated if a user-defined threshold is reached, e.g. the correlation  $r$  between *Customer* and *\_Sales* is  $r = |0.87|$  and therefore higher than the threshold  $t = 0.6$ . The validated elements are promoted to factors and used as a starting point for the next iteration, e.g. the factor *Customer* is one starting point for the second iteration. In the next iteration, the elements connected to the factor *Customer* are activated and validated, e.g. the element *\_Quality* is newly activated (*SalesTerritory* and *Price* were already activated, see above) and the validation of *\_Quality* is successful.

In this proof of concept, the evaluation phase is done without any changes. The possible adjustments from an expert would be comparable to the adjustments already done during the creation of the knowledge representation (ref. Chapter 3.2).

### 3.4 DSS model output

A DSS model is generated to improve the decision-making ability of the user (Holsapple & Whinston 1996, p.136). It supports the design phase of a decision to propose alternatives or showing the effects of various alternatives (Holsapple & Whinston 1996, p.136). The model is based on the activated as well as validated factors and consists of multiple layers due to the multiple iterations during the exploration phase (ref. Chapter 3.3). Each layer adds more detailed explanations for the analyzed effect by including more and more causes (ref. Figure 3).



**Figure 3:** The resulting DSS model for a specific product subcategory and a specific sales territory with multiple layers

The first layer (ref. Layer 1 in Figure 3) represents the initial DSS model including only the scenario without any explanations. The second layer (ref. Layer 2 in Figure 3) explains the effect of *dropping sales* with a decreasing *product quality* by adding the activated and validated factors from the first iteration of the exploration phase to the model. The third layer (ref. Layer 3 in Figure 3) extends the model with activated and validated factors from the second iteration. The decreasing *product quality* is caused by a change in the *working shifts*. This DSS model can be used to simulate potential solutions to the problem, e.g. the cost of the number of employees who would have to be hired to compensate the change in the working shifts could be compared to a possible increase of the service quality.

#### 4. Conclusions and further steps

The feasibility of this concept for a knowledge-based cause-effect analysis was proven by manually applying the concept to an example data set. An automatic extraction of knowledge about cause-effect relationships from database structures, ETL processes and DSS models for the integration into a unified knowledge base is envisioned in the concept (ref. Chapter 2). The automatic creation of the knowledge base could be done for a specific environment (e.g. in the context of a company) and then extended with domain-specific knowledge from experts (ref. Chapter 2). These experts require a user-friendly interface to adjust the knowledge representation and add domain-specific knowledge. This proof of concept also showed, that there are many different parameters during the knowledge reasoning (e.g. the activation strategy, the validation calculation). Because of a very complex architecture and environment, a prototype of the knowledge-based cause-effect analysis will be developed to evaluate the effectiveness of these parameters. This will be done by using a component-based architecture to easily exchange components and compare the results of using different components for the same step (similar to Markus et al. 2002, pp.202–205 for emergent knowledge processes). The research is concluded by evaluating the prototype concerning the support during the design and choice phases of decisions (Holsapple & Whinston 1996, p.136). This is done by demonstrating the prototype as a realizable and valid solution to predefined situations (Vaishnavi & Kuechler, Jr. 2007, p.160). Additionally, an experiment could be used to evaluate the prototype (Boudreau et al. 2001, p.3), e.g. with real-life data in a selected company or in a laboratory setting.

## References

- Benjamins, A., 2014. Knowledge-based cause-effect analysis for context-driven decision support. In *DSS 2.0 – Supporting Decision Making with New Technologies, Supplemental Proceedings*. Paris: IFIP Working Group 8.3 digital publications. Available at: <http://dss20conference.files.wordpress.com/2014/05/benjamins.pdf> [Accessed June 18, 2014].
- Boudreau, M.-C., Gefen, D. & Straub, D.W., 2001. Validation in Information Systems Research: A State-of-the-Art Assessment. *MIS Quarterly*, 25(1), pp.1–16.
- Brachman, R.J., Levesque, H.J. & Pagnucco, M., 2004. *Knowledge representation and reasoning*, Amsterdam; Boston: Morgan Kaufmann.
- Chen, H., Chiang, R.H.L. & Storey, V.C., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36(4), pp.1165–1188.
- Davenport, T.H. & Harris, J.G., 2007. *Competing on Analytics: The New Science of Winning*, Boston, Massachusetts: Harvard Business School Press.
- Gómez-Pérez, A., Fernández-López, M. & Corcho, O., 2003. *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*, London: Springer.
- Holsapple, C.W. & Whinston, A.B., 1996. *Decision support systems: A knowledge-based approach*, St. Paul, Minnesota: West Publishing Company.
- Markus, M.L., Majchrzak, A. & Gasser, L., 2002. A Design Theory for Systems That Support Emergent Knowledge Processes. *MIS Quarterly*, 26(3), pp.179–212.
- Microsoft, 2014. *Microsoft SQL Server Product Samples Database*. Available at: <https://msftdbprodsamples.codeplex.com> [Accessed March 26, 2014].
- Mintzberg, H., Raisinghani, D. & Theoret, A., 1976. The structure of “unstructured” decision processes. *Administrative science quarterly*, 21(2), pp.246–275.
- Newell, A. & Simon, H.A., 1972. *Human problem solving* 6th printing., Englewood Cliffs, New Jersey: Prentice-Hall.
- Simon, H.A., 1977. *The new science of management decision* Revised Edition., Englewood Cliffs, New Jersey: Prentice-Hall.
- Vaishnavi, V.K. & Kuechler, Jr., W., 2007. *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*, Boca Raton, Florida: Auerbach Publications.
- Vercellis, C., 2009. *Business intelligence: Data Mining and Optimization for Decision Making*, Chichester, United Kingdom: John Wiley & Sons.
- W3C OWL Working Group, 2009. *OWL 2 Web Ontology Language Document Overview*. Available at: <http://www.w3.org/TR/owl2-overview> [Accessed March 26, 2014].

# A Simulation Study of Operational Availability of Canadian Armed Forces Equipment under Different Maintenance Regimes

Abdeslem Boukhtouta and Ahmed Ghanmi

Defence Research and Development Canada, Centre for Operational Research and Analysis (DRDC-CORA), Ottawa, Canada

[Abdeslem.boukhtouta@drdc-rddc.gc.ca](mailto:Abdeslem.boukhtouta@drdc-rddc.gc.ca)

[Ahmed.Ghanmi@drdc-rddc.gc.ca](mailto:Ahmed.Ghanmi@drdc-rddc.gc.ca)

**Abstract:** This paper presents a study on operational availability (OA) of Canadian Armed Forces (CAF) equipment. An OA conceptual framework for CAF equipment was developed and implemented in a discrete event simulation model to assess the impact of different maintenance regimes (planned, non-planned, and condition based). The basic premise of the condition based maintenance is that maintenance is conducted on equipment based on evidence of need, rather than any time or usage schedule. In the simulation, activity durations and maintenance decisions were represented by probability distributions. Preliminary results indicated that the OA framework captures general trends of CAF land equipment availability and that the maintenance regimes, particularly the condition based maintenance, would have a great impact on OA. A sensitivity analysis was also conducted to examine options for improving OA of military equipment.

**Keywords:** Operational availability, simulation, condition-based maintenance, corrective maintenance, scheduled maintenance

---

## 1. Introduction

A number of culminating pressures on the Canadian Armed Forces (CAF) equipment, particularly the land equipment, is challenging the CAF ability to meet their operational commitments. Indeed, reports indicated that land equipment availability is at its lowest level in ten years. If not addressed, the deficiencies in the Canadian Army equipment may impact its ability to conduct future operations (See Commander Canadian Army, 2013). Enhancing operational availability (OA) of land systems and reducing the total cost of equipment ownership are high priorities for CAF.

The problem of military systems availability is highly complex, with multiple interdependencies and external influencers (Boukhtouta et al. 2012). Military systems availability is a fundamental requirement for conducting operations. Many logistics and support factors contribute to achieving levels of equipment availability for both training fleets and those deployed to operations. Studies (simulations, theoretical, etc.) are required to identify and understand the relative contributions of these factors and activities, and how they can be employed within the supply chain and the logistics support network to achieve desired levels of OA within the limited Defence budgets (Commander Canadian Army 2013). This paper examines OA of CAF land equipment and focuses on the impact of different maintenance regimes on OA. A high level generic framework, capturing the most important factors influencing the OA, was developed. The framework considers several maintenance regimes, such as Corrective Maintenance (CM), Scheduled Preventive Maintenance (SPM), and Condition-Based Maintenance (CBM) and was implemented in discrete event simulation model using ARENA software. In the simulation, activity durations (repair time, failure time, supply delay, etc.) and maintenance decisions were represented by probability distributions and the expected OA was calculated for a large number of simulation runs.

The paper is organized as follows. Section 2 presents a summary of the literature review related the study presented in this paper. Section 3 describes the OA conceptual framework of CAF equipment. Section 4 presents the simulation model and the impact of different maintenance regimes on OA. Concluding remarks and future work are given in section 5.

## 2. Literature review

Availability of military systems is of major concern for military planners since different factors can critically affect the operational effectiveness during military operations (See US DoD 2014 and Lie 1977). Different papers addressing the OA of military systems from different perspectives have been published in the literature. However and in our knowledge there is no discrete event simulation study of the OA of military systems under different maintenance regimes. We discuss below the studies pertaining to the current paper.



A methodology for estimating OA of military systems is proposed by Gary (2008). The equations and methodologies given in this paper describe the most common techniques, to determine the OA, as well as their limitations and shortcomings. From another side, an investigation study on the effects of the prognostics capability on OA of military land systems has been conducted in Koehn et al. (2005). It is also shown in that paper that OA of military vehicles is significantly affected by Administrative Logistics Delay Time (ALDT) and repair times. Koehn et al. recommended in their paper to use prognostics approaches which allow the ALDT reduction and OA improvement by anticipating failure and preparing the necessary replacement parts. Mi (1998) compared system availability time interval measures of a single-unit system based on stochastic orderings and classifications of lifetime distributions. Murdock (1995) used the renewal theory to develop an availability model over a finite time horizon for a continuously demanded component. He showed that the optimal age replacement period in an infinite time horizon does not maximize average availability. The result of Murdock (1995) study is very useful for lifecycle maintenance planning. A discrete event simulation model using Monte Carlo methods is presented by Sadananda and Srinivasan (2012) to estimate the availability of military systems but not under different maintenance regimes. A system engineering approach is used in this paper and the availability results obtained may also be used for long-term procurement decisions and strategic planning.

### **3. OA Conceptual Framework**

Figure 1 depicts the high level OA concept framework of the CAF Land system. The CAF Land system consists, for the purpose of this study, of fleets of vehicles, which partake in missions (tasks). Between tasks, vehicles are non-operating and awaiting their next task. While non-operating, a vehicle could be down for maintenance (repair, upkeep, or checkup), on standby, or simply available and ready for their next task. There are two main categories of maintenance operations:

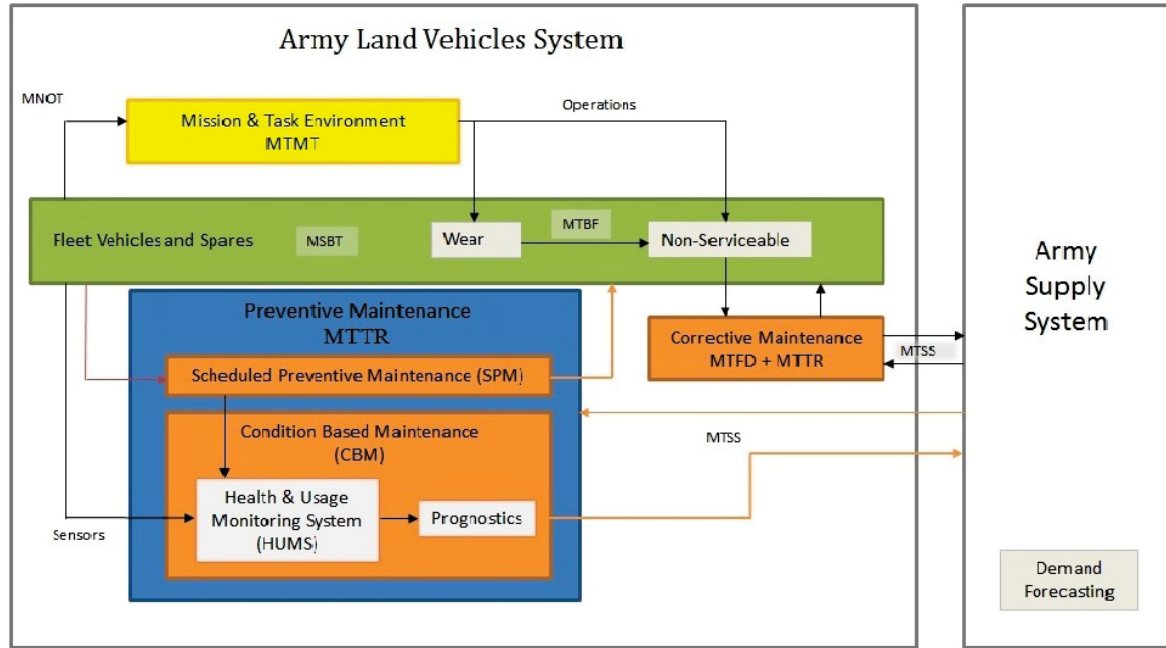
- *Corrective Maintenance (CM)*: CM deals with repair of system faults and requires diagnostics for fault identification. CM is triggered by system failures and the vehicle remains down until repaired.
- *Preventive maintenance (PM)* : two kind of preventive maintenance are considered,
  - SPM: SPM includes upgrades and checkups to decrease likelihood of failure. SPM is triggered via a maintenance tracking system that follows the rules established by the manufacturer. The vehicle is down for SPM when required spares are available only.
  - CBM: CBM is triggered by a decision-making system that uses health and usage monitoring system (HUMS) data to prognose what is most likely to fail, based on current state, if not maintained within a certain time. Similar to SPM, the vehicle is down for CBM when required spares are available only.

All types of maintenance consume spare parts. If spares are not available at the maintenance site, they must be ordered from the Supply System. Demand Forecasting technologies can be used to reduce the Time to Supply Spares (TSS). The prefix M is used in this paper to designate the mean (eg. MTSS, is the Mean Time to Supply Spares).

A vehicle undergoing CM is considered unavailable (down) while spare parts are being shipped to maintenance site. This is referred as Time to Supply Spare (TSS). The time required for CM can be subdivided into Time for Failures diagnostics (TFD), TSS, and Time To Repair (TTR). The latter includes an operational checkout performed to ensure the vehicle is once more serviceable. If the check passes, the vehicle goes on standby (the acronym SBT – Standby Time is used for this time), otherwise the CM must restart with a new attempt at diagnostics. Repair times are typically much smaller than TSS. However, for CBM, TSS is eliminated, and there is no diagnostics time: down time is approximately equal to repair time.

As indicted above, Figure 1 depicts most important concepts (boxes) and relationships (lines). The system could be available (green), in-use (yellow), and unavailable (orange). The Fleet Vehicles box is the main component, representing the fleet of vehicles in standby. On average a vehicle is in a standby state for the mean standby time (MSBT). Vehicles undergo wear as a result of the Mission and Task Environment box, representing the usage of vehicles on mission tasks. Tasks last on average a time MTMT (the mean time for mission tasks). A vehicle may become non-serviceable as a result of wear-induced or combat induced failures,

both experienced during mission tasks. Failures manifest on average every MTBF (mean time between failures) days.



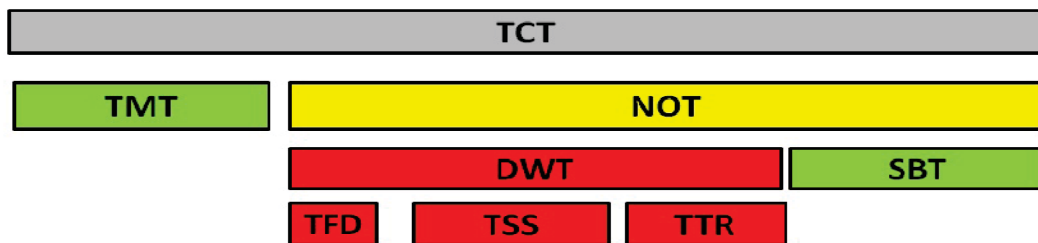
**Figure 1:** High level Conceptual Framework for Analysis of OA of Land Systems

The Preventive Maintenance box is the actual maintenance activity resulting after either SPM or CBM has been triggered and spares are available (no TFD). For CBM, data from the vehicle sensors are fed into the HUMS.

Figure 2 depicts a graphical representation of a given Task Cycle Time (TCT) for a vehicle; each cycle has two times: Time on Missions Tasks (TMT) and Non-operating Time (NOT). The NOT can be decomposed into DWT (Downtime) and SBT (Time on Standby). The DWT is the sum of TFD, TSS, and TTR. Using these notations, the uptime of a vehicle is TMT + SBT whereas its downtime is DWT, which can be approximated by the sum of TSS and TTR (i.e. TFD is treated as negligible in comparison with TSS and TTR). The OA of a vehicle is given by (see Gary 2008):

$$OA = \frac{\text{uptime}}{\text{uptime} + \text{downtime}} \quad (1)$$

$$= \frac{(TMT + SBT)}{(TMT + SBT + TSS + TTR)}$$



**Figure 2:** Time Subdivision Hierarchy for Vehicle OA Measurements

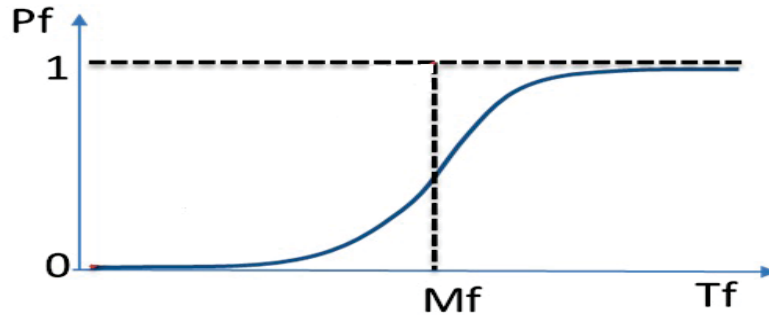
Each of those time durations is treated as a stochastic variable.

#### 4. Simulation study

Only one vehicle type is supported, with one spare part type, and one task type. The sensor variable for CBM is tire thickness. Tire thickness can be interpreted as a metaphor for any sort of “system wear”.

All probability distributions are labelled  $P_x$  where  $x$  is a suffix indicating which component it represents:  $x=t$  for tasks,  $n$  for non-operating,  $r$  for repair, and  $s$  for supply. Each distribution has a mean  $M_x$  and standard deviations  $S_x$ .

The probability of failure  $p_f$  has been approximated by a linear function in Schoenborn, et al. (2014) however  $P_f$  is implemented in the current study as a sigmoid function of the time spent on task ( $T_f$ ) with a mean  $M_f$  (the center of the profile). Thus, the probability density of  $P_f$  follows an approximately Gaussian distribution with mean  $M_f$  (see Figure 3).

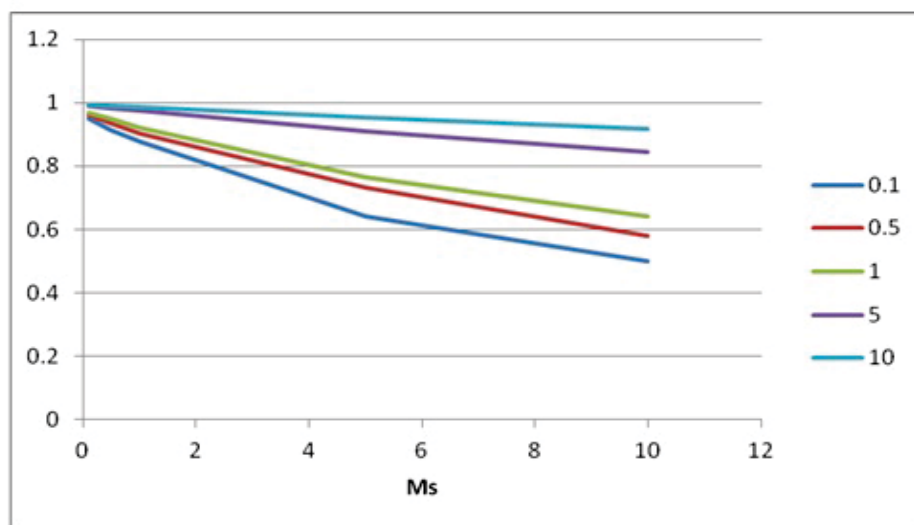


**Figure 1:** Probability of failure function ( $P_f$ )

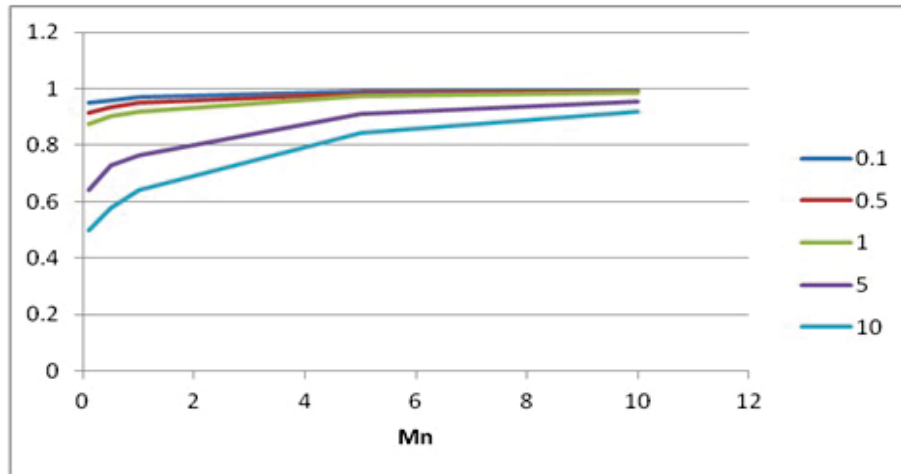
The study presented in this paper is a work in progress and we describe below the preliminary simulation results.

#### 4.1 Corrective Maintenance (CM) OA versus MTSS (or $M_s$ ) and ( $M_n$ )

Figure 4 shows the OA variation versus  $M_s$  for different  $M_n$  values (0.1, 0.5, 1, 5 and 10). Figure 5 shows the OA variation versus  $M_n$  for different  $M_s$  values (0.1, 0.5, 1, 5 and 10). The graphs follow the expected trend that OA significantly decreases as the MTSS increases. The actual value of OA is closely related to the value of  $M_f$ : an  $M_f$  of 20, for an  $M_t$  of 1, implies that on average a failure will happen before 10 task cycles half the time, and after 10 cycles half the time, i.e. about 1/10 cycles will have a failure, which should lead to an OA of approximately 0.9 if the downtime is similar to task duration. This case for  $M_s = 0.1$ , 0.5 and  $M_n = 0.1$  and 0.5. As  $M_s$  increases, the failure rate remains constant but the impact on uptime becomes more important: at  $M_s = 10$ , there is at least a 10-day downtime (since  $M_r > 0$ ), so every 10 cycles of 1 day on average there is 10 days of downtime, leading to an availability of about 0.5. This is supported by the results as shown in Figure 5.



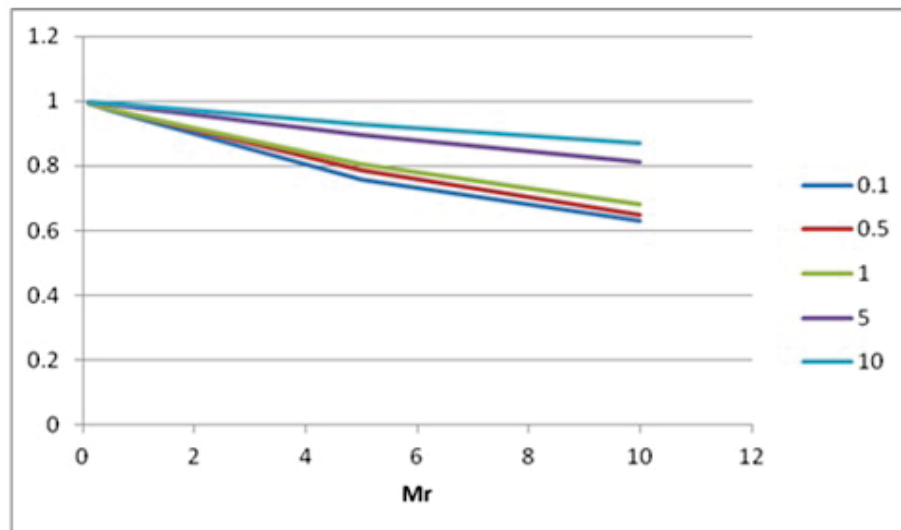
**Figure 2:** CM OA versus  $M_s$  for different  $M_n$  values



**Figure 3:** CM OA versus Mn for different Ms values

### 3.2 CBM OA versus MTTR (or Mr) and MTSS (or Mn)

The results presented in this subsection are specific to CBM when CM never triggered. Figure 6 shows CBM OA variation versus Mr for different Mn values (0.1, 0.5, 1, 5, and 10 in days) however Figure 7 shows CBM OA variation versus Mn for different Mr values (0.1, 0.5, 1, 5, and 10 in days). In a CBM regime without possibility of CM, the supply time is invisible: the vehicle remains available for tasks while the spares are shipped from time of CBM trigger until arrival at maintenance shop. The wear rate is such that CBM will be required on average after 5 tasks (5 days) but will occur only after  $M_s/M_t = 10$  tasks, the time for the spare to arrive, totaling 15 days. At that point, the vehicle will be taken off line for maintenance for Mr days. For Mn very small such that vehicle always as busy as can be, Mr=5 days, which implies an OA of about 75%, indeed observed. For larger Mn, the impact of maintenance on availability is smaller, leading to curves with overall higher OA. This is supported by the results as shown in Figure 7.



**Figure 4:** CBM OA versus Mr for different Mn values

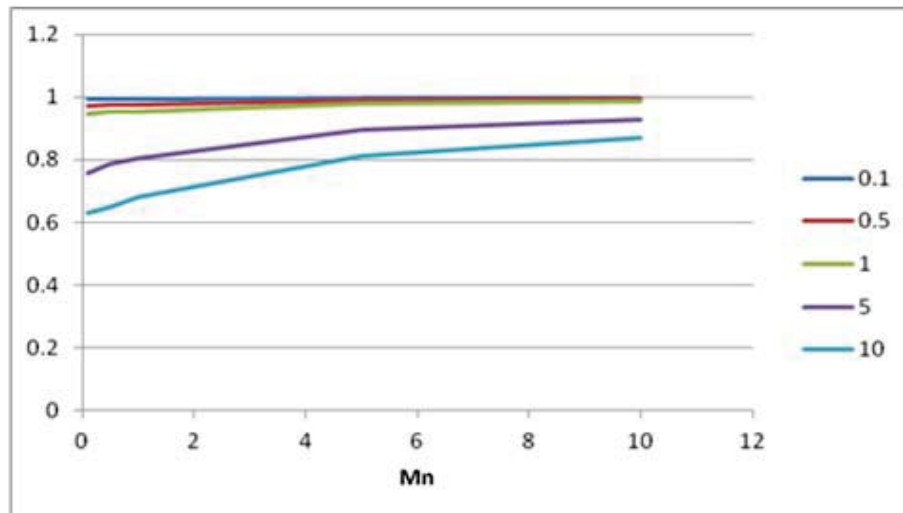


Figure 5: CBM OA versus Mn for different Ms values

## 5. Conclusion

The objective of this paper was to study the OA of CAF Land systems for several maintenance regimes: namely CM, SPM, and CBM. A subset of this simulation study is presented in this paper. Simulations were run with various probability distributions for the various process durations such as task duration, time between failure, repair time, supply delay and non-operating time. The results indicate that the framework and the implementation of its related simulation model properly capture the important trends of OA for a system such as CAF systems for both the CM and CBM regimes.

## References

- Boukhtouta A., Fox C., Fry D., Zhang L., (2012), Literature Review and National Perspectives on Operational Availability and Sparing Analysis, Technical report, LND TTCP AG4, TR-LND-AG5-03-2012
- Commander Canadian Army, (2013) Canadian Army Equipment Readiness Strategy, Canadian National Defence, National Defence Headquarters, Ottawa, 27 May 2013
- Gary A. P. (2008) Methodology for Estimation of operational Availability as applied to Military Systems, ITEA journal, 29, pp. 420-428
- Koehn P., Macheret Y. and Sparrow D. (2005), Improving Reliability and Operational Availability of Military Systems, IEEE Aerospace Conference, Big Sky, Montana, March 5-12, 11 pages.
- Lie, C.H., Hwang, C.L., and F.A. Tillman (1977), Availability of Maintained Systems: A State-of-the-Art Survey, AIIE Transactions, 9(3), 247-259.
- Mi J. (1998), Some Comparison Results of System Availability, Naval Research Logistics, 45, 205-218.
- Murdock, Jr., W.P. (1995), Component Availability for an Age Replacement Preventive Maintenance Policy, Ph.D. Dissertation, Department of Industrial and Systems Engineering, Virginia Tech.
- Sadananda, K. Upadhyay, Srinivasan N.K., (2012), Availability estimation using simulation for military systems, International Journal of Quality & Reliability Management, Vol. 29 Iss: 8, pp.937 – 952.
- Schoenborn, et al., (2014) Conceptual Framework Development for the Operational Availability of the Canadian Army Land Systems, CAE Integrated Enterprise Solutions – Canada, Document No 5568-001, Contract Report.
- US DoD (Department of Defence), (2005), Guide for Achieving Reliability, Availability, and Maintainability, 266 pages. <http://www.weibull.com/knowledge/milhdbk.htm>







# ICAS 2015

For further information contact  
[info@academic-conferences.org](mailto:info@academic-conferences.org)  
or telephone  
+44-(0)-118-972-4148

